

Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) EP 1 349 088 A2

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
01.10.2003 Bulletin 2003/40

(51) Int Cl.7: G06F 17/30, G06F 11/14

(21) Application number: 03251702.1

(22) Date of filing: 19.03.2003

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR
HU IE IT LI LU MC NL PT RO SE SI SK TR
Designated Extension States:
AL LT LV MK

- Manley, Stephen L.
London SW5 9JZ (GB)
- Owara, Shane S.
Mountain View, California 94040 (US)
- Kleiman, Steven R.
Sunnyvale, California 94089 (US)

(30) Priority: 19.03.2002 US 100950
19.03.2002 US 100967

(71) Applicant: Network Appliance, Inc.
Sunnyvale, California 94089 (US)

(74) Representative: Collins, John David et al
Marks & Clerk,
57-60 Lincoln's Inn Fields
London WC2A 3LS (GB)

(72) Inventors:
• Federwisch, Michael L.
San Jose, California 95125 (US)

(54) System and method for determining and transmitting changes in snapshots

(57) A system and method for remote asynchronous replication or mirroring of changes in a source file system snapshot in a destination replica file system using a scan (via a scanner) of the blocks that make up two versions of a snapshot of the source file system, which identifies changed blocks in the respective snapshot files based upon differences in volume block numbers identified in a scan of the logical file block index of each snapshot. Trees of blocks associated with the files are traversed, bypassing unchanged pointers between versions and walking down to identify the changes in the hierarchy of the tree. These changes are transmitted to the destination mirror or replicated snapshot. This technique allows regular files, directories, inodes and any

other hierarchical structure to be efficiently scanned to determine differences between versions thereof. The changes in the files and directories are transmitted over the network for update of the replicated destination snapshot in an asynchronous (lazy write) manner. The changes are described in an extensible, system-independent data stream format layered under a network transport protocol. At the destination, source changes are used to update the destination snapshot. Any deleted or modified inodes already on the destination are moved to a temporary or "purgatory" directory and, if re-used, are relinked to the rebuilt replicated snapshot directory. The source file system snapshots can be representative of a volume sub-organization, such as a qtree.

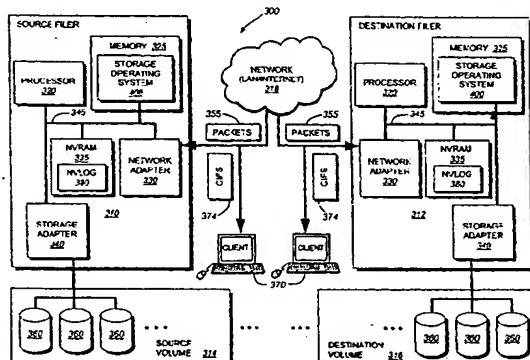


FIG. 3

EP 1 349 088 A2

Description

FIELD OF THE INVENTION

[0001] This invention relates to storage of data using file servers and more particularly to mirroring or replication of stored data in remote storage locations over a network.

BACKGROUND OF THE INVENTION

[0002] A file server is a computer that provides file service relating to the organization of information on storage devices, such as disks. The file server or *filer* includes a storage operating system that implements a file system to logically organize the information as a hierarchical structure of directories and files on the disks. Each "on-disk" file may be implemented as a set of data structures, e.g., disk blocks, configured to store information. A directory, on the other hand, may be implemented as a specially formatted file in which information about other files and directories are stored.

[0003] A filer may be further configured to operate according to a client/server model of information delivery to thereby allow many clients to access files stored on a server, e.g., the filer. In this model, the client may comprise an application, such as a database application, executing on a computer that "connects" to the filer over a direct connection or computer network, such as a point-to-point link, shared local area network (LAN), wide area network (WAN), or virtual private network (VPN) implemented over a public network such as the Internet. Each client may request the services of the file system on the filer by issuing file system protocol messages (in the form of packets) to the filer over the network.

[0004] A common type of file system is a "write in-place" file system, an example of which is the conventional Berkeley fast file system. By "file system" it is meant generally a structuring of data and metadata on a storage device, such as disks, which permits reading/writing of data on those disks. In a write in-place file system, the locations of the data structures, such as inodes and data blocks, on disk are typically fixed. An inode is a data structure used to store information, such as metadata, about a file, whereas the data blocks are structures used to store the actual data for the file. The information contained in an inode may include, e.g., ownership of the file, access permission for the file, size of the file, file type and references to locations on disk of the data blocks for the file. The references to the locations of the file data are provided by pointers in the inode, which may further reference indirect blocks that, in turn, reference the data blocks, depending upon the quantity of data in the file. Changes to the inodes and data blocks are made "in-place" in accordance with the write in-place file system. If an update to a file extends the quantity of data for the file, an additional data block is allocated and

the appropriate inode is updated to reference that data block.

[0005] Another type of file system is a write-anywhere file system that does not overwrite data on disks. If a data block on disk is retrieved (read) from disk into memory and "dirty" with new data, the data block is stored (written) to a new location on disk to thereby optimize write performance. A write-anywhere file system may initially assume an optimal layout such that the data is substantially contiguously arranged on disks. The optimal disk layout results in efficient access operations, particularly for sequential read operations, directed to the disks. A particular example of a write-anywhere file system that is configured to operate on a filer is the Write Anywhere File Layout (WAFL™) file system available from Network Appliance, Inc. of Sunnyvale, California. The WAFL file system is implemented within a microkernel as part of the overall protocol stack of the filer and associated disk storage. This microkernel is supplied as part of Network Appliance's Data ONTAP™ software, residing on the filer, that processes file-service requests from network-attached clients.

[0006] As used herein, the term "storage operating system" generally refers to the computer-executable code operable on a computer that manages data access and may, in the case of a filer, implement file system semantics, such as the Data ONTAP™ storage operating system, implemented as a microkernel, and available from Network Appliance, Inc. of Sunnyvale, California, which implements a Write Anywhere File Layout (WAFL™) file system. The storage operating system can also be implemented as an application program operating over a general-purpose operating system, such as UNIX® or Windows NT®, or as a general-purpose operating system with configurable functionality, which is configured for storage applications as described herein.

[0007] Disk storage is typically implemented as one or more storage "volumes" that comprise physical storage disks, defining an overall logical arrangement of storage space. Currently available filer implementations can serve a large number of discrete volumes (150 or more, for example). Each volume is associated with its own file system and, for purposes hereof, volume and file system shall generally be used synonymously. The disks within a volume are typically organized as one or more groups of Redundant Array of Independent (or *Inexpensive*) Disks (RAID). RAID implementations enhance the reliability/integrity of data storage through the redundant writing of data "stripes" across a given number of physical disks in the RAID group, and the appropriate caching of parity information with respect to the striped data. In the example of a WAFL file system, a RAID 4 implementation is advantageously employed. This implementation specifically entails the striping of data across a group of disks, and separate parity caching within a selected disk of the RAID group. As described herein, a *volume* typically comprises at least one

data disk and one associated parity disk (or possibly data/parity partitions in a single disk) arranged according to a RAID 4, or equivalent high-reliability, implementation.

[0008] In order to improve reliability and facilitate disaster recovery in the event of a failure of a filer, its associated disks or some portion of the storage infrastructure, it is common to "mirror" or replicate some or all of the underlying data and/or the file system that organizes the data. In one example, a mirror is established and stored at a remote site, making it more likely that recovery is possible in the event of a true disaster that may physically damage the main storage location or its infrastructure (e.g. a flood, power outage, act of war, etc.). The mirror is updated at regular intervals, typically set by an administrator, in an effort to catch the most recent changes to the file system. One common form of update involves the use of a "snapshot" process in which the active file system at the storage site, consisting of inodes and blocks, is captured and the "snapshot" is transmitted as a whole, over a network (such as the well-known Internet) to the remote storage site. Generally, a snapshot is an image (typically read-only) of a file system at a point in time, which is stored on the same primary storage device as is the active file system and is accessible by users of the active file system. By "active file system" it is meant the file system to which current input/output operations are being directed. The primary storage device, e.g., a set of disks, stores the active file system, while a secondary storage, e.g. a tape drive, may be utilized to store backups of the active file system. Once snapshotted, the active file system is reestablished, leaving the snapshotted version in place for possible disaster recovery. Each time a snapshot occurs, the old active file system becomes the new snapshot, and the new active file system carries on, recording any new changes. A set number of snapshots may be retained depending upon various time-based and other criteria. The snapshotting process is described in further detail in US2002083037, entitled INSTANT SNAPSHOT by Blake Lewis *et al.* In addition, the native Snapshot™ capabilities of the WAFL file system are further described in *TR3002 File System Design for an NFS File Server Appliance* by David Hitz *et al.*, published by Network Appliance, Inc., and in commonly owned U.S. Patent No. 5,819,292 entitled METHOD FOR MAINTAINING CONSISTENT STATES OF A FILE SYSTEM AND FOR CREATING USER-ACCESSIBLE READ-ONLY COPIES OF A FILE SYSTEM by David Hitz *et al.*

[0009] The complete recopying of the entire file system to a remote (destination) site over a network may be quite inconvenient where the size of the file system is measured in tens or hundreds of gigabytes (even terabytes). This full-backup approach to remote data replication may severely tax the bandwidth of the network and also the processing capabilities of both the destination and source filer. One solution has been to limit the snapshot to only portions of a file system volume that

have experienced changes. Hence, Fig. 1 shows a prior art volume-based mirroring where a source file system 100 is connected to a destination storage site 102 (consisting of a server and attached storage—not shown) via a network link 104. The destination 102 receives periodic snapshot updates at some regular interval set by an administrator. These intervals are chosen based upon a variety of criteria including available bandwidth, importance of the data, frequency of changes and overall volume size.

[0010] In brief summary, the source creates a pair of time-separated snapshots of the volume. These can be created as part of the commit process in which data is committed to non-volatile memory in the filer or by another mechanism. The "new" snapshot 110 is a recent snapshot of the volume's active file system. The "old" snapshot 112 is an older snapshot of the volume, which should match the image of the file system replicated on the destination mirror. Note, that the file server is free to continue work on new file service requests once the new snapshot 112 is made. The new snapshot acts as a checkpoint of activity up to that time rather than an absolute representation of the then-current volume state. A differencer 120 scans the blocks 122 in the old and new snapshots. In particular, the differencer works in a block-by-block fashion, examining the list of blocks in each snapshot to compare which blocks have been allocated. In the case of a write-anywhere system, the block is not reused as long as a snapshot references it, thus a change in data is written to a new block. Where a change is identified (denoted by a presence or absence of an 'X' designating data), a decision process 200, shown in Fig. 2, in the differencer 120 decides whether to transmit the data to the destination 102. The process 200 compares the old and new blocks as follows: (a) Where data is in neither an old nor new block (case 202) as in old/new block pair 130, no data is available to transfer. (b) Where data is in the old block, but not the new (case 204) as in old/new block pair 132, such data has already been transferred, (and any new destination snapshot pointers will ignore it), so the new block state is not transmitted. (c) Where data is present in the both the old block and the new block (case 206) as in the old/new block pair 134, no change has occurred and the block data has already been transferred in a previous snapshot. (d) Finally, where the data is not in the old block, but is in the new block (case 208) as in old/new block pair 136, then a changed data block is transferred over the network to become part of the changed volume snapshot set 140 at the destination as a changed block 142. In the exemplary write-anywhere arrangement, the changed blocks are written to new, unused locations in the storage array. Once all changed blocks are written, a base file system information block, that is the root pointer of the new snapshot, is then committed to the destination. The transmitted file system information block is committed, and updates the overall destination file system by pointing to the changed block

structure in the destination, and replacing the previous file system information block. The changes are at this point committed as the latest incremental update of the destination volume snapshot. This file system accurately represents the "new" snapshot on the source. In time a new "new" snapshot is created from further incremental changes.

[0011] Approaches to volume-based remote mirroring of snapshots are described in detail in commonly owned EP1099165, entitled FILE SYSTEM IMAGE TRANSFER by Steven Kleiman, *et al.* and EP1230598, entitled FILE SYSTEM IMAGE TRANSFER BETWEEN DISSIMILAR FILE SYSTEMS by Steven Kleiman, *et al.*

[0012] This volume-based approach to incremental mirroring from a source to a remote storage destination is effective, but may still be inefficient and time-consuming as it forces an entire volume to be scanned for changes and those changes to be transmitted on a block-by-block basis. In other words, the scan focuses on blocks without regard to any underlying information about the files, inodes and data structures, which the blocks comprise. The destination is organized as a set of volumes so a direct volume-by-volume mapping is established between source and destination. Again, where a volume may contain a terabyte or more of information, the block-by-block approach to scanning and comparing changes may still involve significant processor overhead and associated processing time. Often, there may have been only minor changes in a sub-block beneath the root inode block being scanned. Since a list of all blocks in the volume is being examined, however, the fact that many groupings of blocks (files, inode structures, etc.) are unchanged is not considered. In addition, the increasingly large size and scope of a full volume make it highly desirable to sub-divide the data being mirrored into sub-groups, because some groups are more likely to undergo frequent changes, it may be desirable to update their replicas more often than other, less-frequently changed groups. In addition, it may be desirable to mingle original and replicated (snapshotted) sub-groups in a single volume and migrate certain key data to remote locations without migrating an entire volume. Accordingly, a more sophisticated approach to scanning and identifying changed blocks may be desirable, as well as a sub-organization for the volume that allows for the mirroring of less-than-an-entire volume.

[0013] One such sub-organization of a volume is the well-known qtree. Qtrees, as implemented on an exemplary storage system such as described herein, are subtrees in a volume's file system. One key feature of qtrees is that, given a particular qtree, any file or directory in the system can be quickly tested for membership in that qtree, so they serve as a good way to organize the file system into discrete data sets. The use of qtrees as a source and destination for snapshotted data is desirable.

[0014] When a qtree is snapshot is replicated at the destination, it is typically made available for disaster re-

covery and other uses, such as data distribution. However, the snapshot residing on the destination's active file system may be in the midst of receiving or processing an update from the source snapshot when access by a user or process is desired. A way to allow the snapshot to complete its update without interference is highly desirable. Likewise, when a snapshot must return to an earlier state, a way to efficiently facilitate such a return or "rollback" is desired. A variety of other techniques for manipulating different point in time snapshots may increase the versatility and utility of a snapshot replication mechanism.

[0015] In addition, the speed at which a destination snapshot may be updated is partially depends upon the speed with which change data can be committed from the source to the destination's active file system. Techniques for improving the efficiency of file deletion and modification are also highly desirable.

20 SUMMARY OF THE INVENTION

[0016] One aspect of the present invention overcomes the disadvantages of the prior art by providing a system and method for remote asynchronous replication or mirroring of changes in a source file system snapshot in a destination replica file system using a scan (via a scanner) of the blocks that make up two versions of a snapshot of the source file system, which identifies changed blocks in the respective snapshot files based upon differences in volume block numbers identified in a scan of the logical file block index of each snapshot. Trees of blocks associated with the files are traversed, bypassing unchanged pointers between versions and walking down to identify the changes in the hierarchy of the tree. These changes are transmitted to the destination mirror or replicated snapshot. This technique allows regular files, directories, inodes and any other hierarchical structure to be efficiently scanned to determine differences between versions thereof.

[0017] According to an illustrative embodiment, the source scans, with the scanner, along the index of logical file blocks for each snapshot looking for changed volume block numbers between the two source snapshots. Since disk blocks are always rewritten to new locations on the disk, a difference indicates changes in the underlying inodes of the respective blocks. Using the scanner, unchanged blocks are efficiently overlooked, as their inodes are unchanged. Using an inode picker process, that receives changed blocks from the scanner the source picks out inodes from changed blocks specifically associated with the selected qtree (or other sub-organization of the volume). The picker process looks for versions of inodes that have changed between the two snapshots and picks out the changed version. If inodes are the same, but files have changed (based upon different generation numbers in the inodes) the two versions of the respective inodes are both picked out. The changed versions of the inodes (between the two snap-

shots) are queued and transferred to a set of inode handlers/workers or handlers that resolve the changes in underlying blocks by continuing to scan (with the scanner, again) file offsets down "trees" of the inodes until differences in underlying blocks are identified via their block pointers, as changed inodes in one version will point to different data blocks than those in the other version. Only the *changes* in the trees are transmitted over the network for update of the destination file system in an asynchronous (lazy write) manner. The destination file system is exported read-only to the user. This ensures that only the replicator can alter the state of the replica file system.

[0018] In an illustrative embodiment, a file system-independent format is used to transmit a data stream of change data over the network. This format consists of a set of standalone headers with unique identifiers. Some headers refer to follow-on data and others carry relevant data within their stream. For example, the information relating to any source snapshot deleted files are carried within "deleted files" headers. All directory activity is transmitted first, followed by file data. File data is sent in chunks of varying size, separated by regular headers until an ending header (footer) is provided. At the destination, the format is unpacked and inodes contained therein are transmitted over the network are mapped to a new directory structure. Received file data blocks are written according to their offset in the corresponding destination file. An inode map stores entries which map the source's inodes (files) to the destination's inodes (files). The inode map also contains generation numbers. The tuple of (inode number, generation number) allows the system to create a file handle for fast access to a file. It also allows the system to track changes in which a file is deleted and its inode number is reassigned to a newly created file. To facilitate construction of a new directory tree on the destination, an initial directory stage of the destination mirror process receives source directory information via the format and moves any deleted or moved files to a temporary or "purgatory" directory. The purgatory files which have been moved are hard linked from the purgatory directory to the directories where they have been moved to. Newly created source files are entered into map and built into the directory tree. After the directory tree is built, the transfer of file data begins. Changes to file data from the source are written to the corresponding replica files (as identified by the inode map). When the data stream transfer is complete, the purgatory directory is removed and any unlinked files (including various deleted files) are permanently deleted. In one embodiment, a plurality of discrete source qtrees or other sub-organizations derived from different source volumes can be replicated/ mirrored on a single destination volume.

[0019] Another aspect of this invention overcomes the disadvantages of the prior art, in a system and method for updating a replicated destination file system snapshot with changes in a source file system snapshot, by

facilitating construction of a new directory tree on the destination from source update information using a temporary or "purgatory" directory that allows any modified and deleted files on the destination active file system to be associated with (e.g. moved to) the purgatory directory if and until they are reused. In addition, an inode map is established on the destination that maps source inode numbers to destination inode numbers so as to facilitate building of the destination tree using inode/ generation number *tuples*. The inode map allows resynchronization of the source file system to the destination. The inode map also allows association of two or more destination snapshots to each other based upon their respective maps with the source.

[0020] In an illustrative embodiment, a file system-independent format is used to transmit a data stream of changed file data blocks with respect to a source's base and incremental snapshots. Received file data blocks are written according to their offset in the corresponding destination file. An inode map stores entries which map the source's inodes (files) to the destination's inodes (files). The inode map also contains generation numbers. The tuple of (inode number, generation number) allows the system to create a file handle for fast access to a file. It also allows the system to track changes in which a file is deleted and its inode number is reassigned to a newly created file. To facilitate construction of a new directory tree on the destination, an initial directory stage of the destination mirror process receives source directory information via the format and moves any deleted or moved files to the temporary or "purgatory" directory. The purgatory files which have been moved are hard linked from the purgatory directory to the directories where they have been moved to. Newly created source files are entered into map and built into the directory tree. After the directory tree is built, the transfer of file data begins. Changes to file data from the source are written to the corresponding replica files (as identified by the inode map). When the data stream transfer is complete, the purgatory directory is removed and any unlinked files (including various deleted files) are permanently deleted. In one embodiment, a plurality of discrete source qtrees or other sub-organizations derived from different source volumes can be replicated/ mirrored on a single destination volume.

[0021] In another illustrative embodiment, the replicated file system is, itself snapshotted, thereby creating a first exported snapshot. The first exported snapshot corresponds to a first state. If a disaster or communication breakdown occurs after further modifications or updates of the replicated snapshot, then further modification to the replica file system is halted/frozen and a subsequent second exported snapshot is created from the frozen replica file system representing the second state. The replicated file system can be "rolled back" from the second state to the first state by determining the differences in data between the second state and the first state and then applying those changes to recreate the

first state.

[0022] In yet another illustrative embodiment, the inode map used to map inodes transferred from the source snapshot to inodes in the destination replica/mirror file system is used to resynchronize the source state with the destination state. The destination becomes a new "source" and negotiates the transfer of the inode map to the old "source" now the new "destination." The received old inode map is stored on the source and accessed by a flip procedure that generates a new destination map with N inodes equal to the number of inodes on the new source. The new destination then creates entries from the stored source map for each new destination associated with the new source entry if available. Associated generation numbers are also mapped, thereby providing the needed file access tuple. Any entries on the new source index that lack a new destination are marked as zero entries. The completed flipped inode map allows the new source to update the new destination with its changed data.

[0023] In a related embodiment, two replica/mirror snapshots of the same source can establish a mirror relationship with one another. As in the flip embodiment above, the new "source" (old replica) transfers its inode map to the destination system. The destination system then determines the relationship between the two system's inodes. An "associative" process walks the inode maps at the same time (e.g. concurrently, inode number-by-inode number). For each inode from the original source, the process extracts the "destination inode/generation" from each of the inode maps. It then treats the new source as the appropriate map index for the new inode map. It stores the new source generation number, as well as the destination inode/generation.

BRIEF DESCRIPTION OF THE DRAWINGS

[0024] The above and further advantages of the invention may be better understood by referring to the following description in conjunction with the accompanying drawings in which like reference numerals indicate identical or functionally similar elements:

Fig. 1, already described, is a schematic block diagram of an exemplary remote mirroring of a volume snapshot from a source file server to a destination file server over a network according to a prior implementation;

Fig. 2, already described, is a decision table used by a block differencer of Fig. 1 for determining whether a change in a block is to be transmitted from the source file server to the destination file server according to a prior implementation;

Fig. 3 is a schematic block diagram defining an exemplary network and file server environment including a source file server and a destination file server within which the principles of this invention are implemented;

Fig. 4 is a schematic block diagram of an exemplary storage operating system for use with the file servers of Fig. 3;

Fig. 5 is schematic block diagram of an exemplary file system inode structure;

Fig. 6 is a schematic block diagram of the exemplary file system inode structure of Fig. 5 including a snapshot inode;

Fig. 7 is a schematic block diagram of the exemplary file system inode structure of Fig. 6 after data block has been rewritten;

Fig. 8 is a schematic block diagram of an exemplary operation of the snapshot mirroring process at the source;

Fig. 8A is a decision table used in connection with an inode picker process in the snapshot mirroring process of Fig. 8;

Fig. 8B is a more detailed schematic diagram of an exemplary base snapshot and incremental snapshot block illustrating the inode picker process of Fig. 8A;

Fig. 9 is a schematic block diagram of an exemplary operation of an inode worker used in connection with the snapshot mirroring process of Fig. 8;

Fig. 10 is a schematic block diagram of the source file server snapshot mirroring process, the destination snapshot mirroring process, and the communication link between them;

Fig. 11 is a schematic block diagram of a standalone header structure for use in the data stream transmission format between the source and the destination according to an illustrative embodiment;

Fig. 12 is a schematic block diagram of the data stream transmission format between the source and the destination according to an illustrative embodiment;

Fig. 13 is a schematic block diagram of the stages of the snapshot mirroring process on the destination;

Fig. 14 is a schematic block diagram of a generalized inode map for mapping source inodes to the destination snapshot mirror according to an illustrative embodiment;

Fig. 15 is a highly schematic diagram of the population of data files in the destination snapshot mirror at mapped offsets with respect to source data files;

Fig. 16 is a flow diagram of a snapshot rollback procedure according to an illustrative embodiment; and

Fig. 17 is a flow diagram of a inode map flipping procedure for rolling back or resynchronizing the source file system to a state of the destination mirror snapshot according to an illustrative embodiment;

Fig. 18 is a schematic block diagram of an exemplary inode map residing on the destination for use in the flipping procedure of Fig. 17;

Fig. 19 is a schematic block diagram of an exemplary inode map constructed on the old source (new destination) according to the flipping procedure of

Fig. 18;

Fig. 20 is a schematic block diagram of a generalized inode map association process according to an illustrative embodiment.

DETAILED DESCRIPTION OF AN ILLUSTRATIVE EMBODIMENT

A. Network and File Server Environment

[0025] By way of further background, Fig. 3 is a schematic block diagram of a storage system environment 300 that includes a pair of interconnected file servers including a source file server 310 and a destination file server 312 that may each be advantageously used with the present invention. For the purposes of this description, the source file server is a networked computer that manages storage one or more source volumes 314, each having an array of storage disks 360 (described further below). Likewise, the destination filer 312 manages one or more destination volumes 316, also comprising arrays of disks 360. The source and destination file servers or "filers" are linked via a network 318 that can comprise a local or wide area network, such as the well-known Internet. An appropriate network adapter 330 residing in each filer 310, 312 facilitates communication over the network 318. Also for the purposes of this description, like components in each of the source and destination filer, 310 and 312 respectively, are described with like reference numerals. As used herein, the term "source" can be broadly defined as a location from which the subject data of this invention travels and the term "destination" can be defined as the location to which the data travels. While a source filer and a destination filer, connected by a network, is a particular example of a source and destination used herein, a source and destination could be computers/filers linked via a direct link, or via loopback (a "networking" arrangement internal to a single computer for transmitting a data stream between local source and local destination), in which case the source and the destination are the same filer. As will be described further below, the source and destination are broadly considered to be a source sub-organization of a volume and a destination sub-organization of a volume. Indeed, in at least one special case the source and destination sub-organizations can be the same at different points in time.

[0026] In the particular example of a pair of networked source and destination filers, each filer 310 and 312 can be any type of special-purpose computer (e.g., server) or general-purpose computer, including a standalone computer. The source and destination filers 310, 312 each comprise a processor 320, a memory 325, a network adapter 330 and a storage adapter 340 interconnected by a system bus 345. Each filer 310, 312 also includes a storage operating system 400 (Fig. 4) that implements a file system to logically organize the information as a hierarchical structure of directories and files

on the disks.

[0027] It will be understood to those skilled in the art that the inventive technique described herein may apply to any type of special-purpose computer (e.g., file serving appliance) or general-purpose computer, including a standalone computer, embodied as a storage system. To that end, the filers 310 and 312 can each be broadly, and alternatively, referred to as storage systems. Moreover, the teachings of this invention can be adapted to a variety of storage system architectures including, but not limited to, a network-attached storage environment, a storage area network and disk assembly directly-attached to a client/host computer. The term "storage system" should, therefore, be taken broadly to include such arrangements.

[0028] In the illustrative embodiment, the memory 325 comprises storage locations that are addressable by the processor and adapters for storing software program code. The memory comprises a form of random access memory (RAM) that is generally cleared by a power cycle or other reboot operation (i.e., it is "volatile" memory). The processor and adapters may, in turn, comprise processing elements and/or logic circuitry configured to execute the software code and manipulate the data structures. The operating system 400, portions of which are typically resident in memory and executed by the processing elements, functionally organizes the filer by, *inter alia*, invoking storage operations in support of a file service implemented by the filer. It will be apparent to those skilled in the art that other processing and memory means, including various computer readable media, may be used for storing and executing program instructions pertaining to the inventive technique described herein.

[0029] The network adapter 330 comprises the mechanical, electrical and signaling circuitry needed to connect each filer 310, 312 to the network 318, which may comprise a point-to-point connection or a shared medium, such as a local area network. Moreover the source filer 310 may interact with the destination filer 312 in accordance with a client/server model of information delivery. That is, the client may request the services of the filer, and the filer may return the results of the services requested by the client, by exchanging packets 355 encapsulating, e.g., the TCP/IP protocol or another network protocol format over the network 318.

[0030] The storage adapter 340 cooperates with the operating system 400 (Fig. 4) executing on the filer to access information requested by the client. The information may be stored on the disks 360 that are attached, via the storage adapter 340 to each filer 310, 312 or other node of a storage system as defined herein. The storage adapter 340 includes input/output (I/O) interface circuitry that couples to the disks over an I/O interconnect arrangement, such as a conventional high-performance, Fibre Channel serial link topology. The information is retrieved by the storage adapter and processed by the processor 320 as part of the snapshot procedure,

to be described below, prior to being forwarded over the system bus 345 to the network adapter 330, where the information is formatted into a packet and transmitted to the destination server as also described in detail below.

[0031] Each filer may also be interconnected with one or more clients 370 via the network adapter 330. The clients transmit requests for file service to the source and destination filers 310, 312, respectively, and receive responses to the requests over a LAN or other network (318). Data is transferred between the client and the respective filer 310, 312 using data packets 374 defined as an encapsulation of the Common Internet File System (CIFS) protocol or another appropriate protocol such as NFS.

[0032] In one exemplary filer implementation, each filer 310, 312 can include a nonvolatile random access memory (NVRAM) 335 that provides fault-tolerant backup of data, enabling the integrity of filer transactions to survive a service interruption based upon a power failure, or other fault. The size of the NVRAM depends in part upon its implementation and function in the file server. It is typically sized sufficiently to log a certain time-based chunk of transactions (for example, several seconds worth). The NVRAM is filled, in parallel with the buffer cache, after each client request is completed, but before the result of the request is returned to the requesting client.

[0033] In an illustrative embodiment, the disks 360 are arranged into a plurality of volumes (for example, source volumes 314 and destination volumes 316), in which each volume has a file system associated therewith. The volumes each include one or more disks 360. In one embodiment, the physical disks 360 are configured into RAID groups so that some disks store striped data and some disks store separate parity for the data, in accordance with a preferred RAID 4 configuration. However, other configurations (e.g. RAID 5 having distributed parity across stripes) are also contemplated. In this embodiment, a minimum of one parity disk and one data disk is employed. However, a typical implementation may include three data and one parity disk per RAID group, and a multiplicity of RAID groups per volume.

B. Storage Operating System

[0034] To facilitate generalized access to the disks 360, the storage operating system 400 (Fig. 4) implements a write-anywhere file system that logically organizes the information as a hierarchical structure of directories and files on the disks. Each "on-disk" file may be implemented as a set of disk blocks configured to store information, such as data, whereas the directory may be implemented as a specially formatted file in which references to other files and directories are stored. As noted and defined above, in the illustrative embodiment described herein, the storage operating system is the NetApp® Data ONTAP™ operating system available from Network Appliance, Inc., of Sunnyvale, CA that im-

plements the Write Anywhere File Layout (WAFL™) file system. It is expressly contemplated that any appropriate file system can be used, and as such, where the term "WAFL" is employed, it should be taken broadly to refer to any file system that is otherwise adaptable to the teachings of this invention.

[0035] The organization of the preferred storage operating system for each of the exemplary filers is now described briefly. However, it is expressly contemplated that the principles of this invention can be implemented using a variety of alternate storage operating system architectures. In addition, the particular functions implemented on each of the source and destination filers 310, 312 may vary. As shown in Fig. 4, the exemplary storage operating system 400 comprises a series of software layers, including a media access layer 405 of network drivers (e.g., an Ethernet driver). The operating system further includes network protocol layers, such as the Internet Protocol (IP) layer 410 and its supporting transport mechanisms, the Transport Control Protocol (TCP) layer 415 and the User Datagram Protocol (UDP) layer 420. A file system protocol layer provides multi-protocol data access and, to that end, includes support for the CIFS protocol 425, the NFS protocol 430 and the Hypertext Transfer Protocol (HTTP) protocol 435. In addition, the storage operating system 400 includes a disk storage layer 440 that implements a disk storage protocol, such as a RAID protocol, and a disk driver layer 445, that implements a disk control protocol such as the small computer system interface (SCSI).

[0036] Bridging the disk software layers with the network and file system protocol layers is a file system layer 450 of the storage operating system 400. Generally, the layer 450 implements a file system having an on-disk format representation that is block-based using, e.g., 4-kilobyte (KB) data blocks and using inodes to describe the files. In response to transaction requests, the file system generates operations to load (retrieve) the requested data from volumes if it is not resident "in-core", i.e., in the filer's memory 325. If the information is not in memory, the file system layer 450 indexes into the inode file using the inode number to access an appropriate entry and retrieve a volume block number. The file system layer 450 then passes the volume block number to the disk storage (RAID) layer 440, which maps that volume block number to a disk block number and sends the latter to an appropriate driver (for example, an encapsulation of SCSI implemented on a fibre channel disk interconnection) of the disk driver layer 445. The disk driver accesses the disk block number from volumes and loads the requested data in memory 325 for processing by the filer 310, 312. Upon completion of the request, the filer (and storage operating system) returns a reply, e.g., a conventional acknowledgement packet 374 defined by the CIFS specification, to the client 370 over the respective network connection 372.

[0037] It should be noted that the software "path" 470 through the storage operating system layers described

above needed to perform data storage access for the client request received at the filer may alternatively be implemented in hardware or a combination of hardware and software. That is, in an alternate embodiment of the invention, the storage access request data path 470 may be implemented as logic circuitry embodied within a field programmable gate array (FPGA) or an application specific integrated circuit (ASIC). This type of hardware implementation increases the performance of the file service provided by filer 310, 312 in response to a file system request packet 374 issued by the client 370. [0038] Overlying the file system layer 450 is the snapshot mirroring (or replication) application 490 in accordance with an illustrative embodiment of this invention. This application, as described in detail below, is responsible (on the source side) for the scanning and transmission of changes in the snapshot from the source filer 310 to the destination filer 312 over the network. This application is responsible (on the destination side) for the generation of the updated mirror snapshot from received information. Hence, the particular function of the source and destination applications are different, and are described as such below. The snapshot mirroring application 490 operates outside of the normal request path 470 as shown by the direct links 492 and 494 to the TCP/IP layers 415, 410 and the file system snapshot mechanism (480). Notably, the application interacts with the file system layer to gain knowledge of files so it is able to use a file-based data structure (inode files, in particular) to replicate source snapshots at the destination.

C. Snapshot Procedures

[0039] The inherent Snapshot™ capabilities of the exemplary WAFL file system are further described in *TR3002 File System Design for an NFS File Server Appliance* by David Hitz *et al.*, published by Network Appliance, Inc. Note, "Snapshot" is a trademark of Network Appliance, Inc. It is used for purposes of this patent to designate a persistent consistency point (CP) image. A persistent consistency point image (PCPI) is a point-in-time representation of the storage system, and more particularly, of the active file system, stored on a storage device (e.g., on disk) or in other persistent memory and having a name or other unique identifiers that distinguishes it from other PCPIs taken at other points in time. A PCPI can also include other information (metadata) about the active file system at the particular point in time for which the image is taken. The terms "PCPI" and "snapshot" shall be used interchangeably through out this patent without derogation of Network Appliance's trademark rights.

[0040] Snapshots are generally created on some regular schedule. This schedule is subject to great variation. In addition, the number of snapshots retained by the filer is highly variable. Under one storage scheme, a number of recent snapshots are stored in succession

(for example, a few days worth of snapshots each taken at four-hour intervals), and a number of older snapshots are retained at increasing time spacings (for example, a number of daily snapshots for the previous week(s) and weekly snapshot for the previous few months). The snapshot is stored on-disk along with the active file system, and is called into the buffer cache of the filer memory as requested by the storage operating system 400 or snapshot mirror application 490 as described further below. However, it is contemplated that a variety of snapshot creation techniques and timing schemes can be implemented within the teachings of this invention.

[0041] An exemplary file system inode structure 500 according to an illustrative embodiment is shown in Fig. 5. The *inode for the inode file* or more generally, the "root" inode 505 contains information describing the inode file 508 associated with a given file system. In this exemplary file system inode structure root inode 505 contains a pointer to the inode file indirect block 510. The inode file indirect block 510 points to one or more inode file direct blocks 512, each containing a set of pointers to inodes 515 that make up the inode file 508. The depicted subject inode file 508 is organized into volume blocks (not separately shown) made up of inodes 515 which, in turn, contain pointers to file data (or "disk") blocks 520A, 520B and 520C. In the diagram, this is simplified to show just the inode itself containing pointers to the file data blocks. Each of the file data blocks 520 (A-C) is adapted to store, in the illustrative embodiment, 4 kilobytes (KB) of data. Note, however, where more than a predetermined number of file data blocks are referenced by an inode (515) one or more indirect blocks 525 (shown in phantom) are used. These indirect blocks point to associated file data blocks (not shown). If an inode (515) points to an indirect block, it cannot also point to a file data block, and *vice versa*.

[0042] When the file system generates a snapshot of a given file system, a snapshot inode is generated as shown in Fig. 6. The snapshot inode 605 is, in essence, a duplicate copy of the root inode 505 of the file system 500. Thus, the exemplary file system structure 600 includes the same inode file indirect block 510, inode file direct block 512, inodes 515 and file data blocks 520 (A-C) as depicted in Fig. 5. When a user modifies a file data block, the file system layer writes the new data block to disk and changes the active file system to point to the newly created block. The file layer does not write new data to blocks which are contained in snapshots.

[0043] Fig. 7 shows an exemplary inode file system structure 700 after a file data block has been modified. In this illustrative example, file data which is stored at disk block 520C is modified. The exemplary WAFL file system writes the modified contents to disk block 520C', which is a new location on disk. Because of this new location, the inode file data which is stored at disk block (515) is rewritten so that it points to block 520C'. This modification causes WAFL to allocate a new disk block (715) for the updated version of the data at 515. Simi-

larly, the inode file indirect block 510 is rewritten to block 710 and direct block 512 is rewritten to block 712, to point to the newly revised inode 715. Thus, after a file data block has been modified the snapshot inode 605 contains a pointer to the original inode file system indirect block 510 which, in turn, contains a link to the inode 515. This inode 515 contains pointers to the original file data blocks 520A, 520B and 520C. However, the newly written inode 715 includes pointers to unmodified file data blocks 520A and 520B. The inode 715 also contains a pointer to the modified file data block 520C' representing the new arrangement of the active file system. A new file system root inode 705 is established representing the new structure 700. Note that metadata in any snapshotted blocks (e.g. blocks 510, 515 and 520C) protects these blocks from being recycled or overwritten until they are released from all snapshots. Thus, while the active file system root 705 points to new blocks 710, 712, 715 and 520C', the old blocks 510, 515 and 520C are retained until the snapshot is fully released.

[0044] In accordance with an illustrative embodiment of this invention the source utilizes two snapshots, a "base" snapshot, which represents the image of the replica file system on the destination, and an "incremental" snapshot, which is the image that the source system intends to replicate to the destination, to perform needed updates of the remote snapshot mirror to the destination. In one example from the standpoint of the source, the incremental snapshot can comprise a most-recent snapshot and the base can comprise a less-recent snapshot, enabling an up-to-date set of changes to be presented to the destination. This procedure shall now be described in greater detail.

D. Remote Mirroring

[0045] Having described the general procedure for deriving a snapshot, the mirroring of snapshot information from the source filer 310 (Fig. 3) to a remote destination filer 312 is described in further detail. As discussed generally above, the transmission of incremental changes in snapshot data based upon a comparison of changed blocks in the whole volume is advantageous in that it transfers only incremental changes in data rather than a complete file system snapshot, thereby allowing updates to be smaller and faster. However, a more efficient and/or versatile procedure for incremental remote update of a destination mirror snapshot is contemplated according to an illustrative embodiment of this invention. Note, as used herein the term "replica snapshot," "replicated snapshot" or "mirror snapshot" shall be taken to also refer generally to the file system on the destination volume that contains the snapshot where appropriate (for example where a *snapshot of a snapshot* is implied).

[0046] As indicated above, it is contemplated that this procedure can take advantage of a sub-organization of a volume known as a qtree. A qtree acts similarly to lim-

its enforced on collections of data by the size of a partition in a traditional Unix® or Windows® file system, but with the flexibility to subsequently change the limit, since qtrees have no connection to a specific range of blocks on a disk. Unlike volumes, which are mapped to particular collections of disks (e.g. RAID groups of n disks) and act more like traditional partitions, a qtree is implemented at a higher level than volumes and can, thus, offer more flexibility. Qtrees are basically an abstraction in the software of the storage operating system. Each volume may, in fact, contain multiple qtrees. The granularity of a qtree can be sized to just as a few kilobytes of storage. Qtree structures can be defined by an appropriate file system administrator or user with proper permission to set such limits.

[0047] Note that the above-described qtree organization is exemplary and the principles herein can be applied to a variety of file system organizations including a whole-volume approach. A qtree is a convenient organization according to the illustrative embodiment, at least in part, because of its available identifier in the inode file.

[0048] Before describing further the process of deriving changes in two source snapshots, from which data is transferred to a destination for replication of the source at the destination, general reference is made again to the file block structures shown in Figs. 5-7. Every data block in a file is mapped to disk block (or volume block). Every disk/volume block is enumerated uniquely with a discrete volume block number (VBN). Each file is represented by a single inode, which contains pointers to these data blocks. These pointers are VBNs—each pointer field in an inode having a VBN in it, whereby a file's data is accessed by loading up the appropriate disk/volume block with a request to the file system (or disk control) layer. When a file's data is altered, a new disk block is allocated to store the changed data. The VBN of this disk block is placed in the pointer field of the inode. A snapshot captures the inode at a point in time, and all the VBN fields in it.

[0049] In order to scale beyond the maximum number of VBN "pointers" in an inode, "indirect blocks" are used. In essence, a disk block is allocated and filled with the VBNs of the data blocks, the inode pointers then point to the indirect block. There can exist several levels of indirect blocks, which can create a large tree structure. Indirect blocks are modified in the same manner as regular data blocks are—every time a VBN in an indirect block changes, a new disk/volume block is allocated for the altered data of the indirect block.

1. Source

[0050] Fig. 8 shows an exemplary pair of snapshot inode files within the source environment 800. In an illustrative embodiment, these represent two snapshots' mode files: the base 810 and incremental 812. Note that these two snapshots were taken at two points in time;

the base represents the current image of the replica, and the incremental represents the image the replica will be updated to. The differences between the two snapshots define which changes are to be derived and committed to the remote replica/mirror. The inode files may each be loaded into the buffer cache of the source file server memory from the on-disk versions thereof using conventional disk access processes as directed by the storage operating system snapshot manager (480 in Fig. 4). In one embodiment, the base and incremental snapshots are loaded in increments as they are worked on by the operating system (rather than all-at-once). Each snapshot inode file 810, 812 is organized into a series of storage blocks 814. In this illustrative example, the base snapshot inode file 810 contains storage blocks denoted by volume (disk) block numbers, 5, 6 and 7, while the incremental snapshot inode file contains exemplary storage blocks having volume block numbers 3, 5, 6 and 8. Within each of the blocks are organized a given number of inodes 816. The volume blocks are indexed in the depicted order based upon their underlying logical file block placement.

[0051] In the example of a write-anywhere file layout, storage blocks are not immediately overwritten or reused. Thus changes in a file comprised of a series of volume blocks will always result in the presence of a new volume block number (newly written-to) that can be detected at the appropriate logical file block offset relative to an old block. The existence of a changed volume block number at a given offset in the index between the base snapshot inode file and incremental snapshot inode file generally indicates that one or more of the underlying inodes and files to which the inodes point have been changed. Note, however, that the system may rely on other indicators of changes in the inodes or pointers—this may be desirable where a write-in-place file system is implemented.

[0052] A scanner 820 searches the index for changed base/incremental inode file snapshot blocks, comparing volume block numbers or another identifier. In the example of Fig. 8, block 4 in the base snapshot inode file 810 now corresponds in the file scan order to block 3 in the incremental snapshot inode file 812. This indicates a change of one or more underlying inodes. In addition, block 7 in the base snapshot inode file appears as block 8 in the incremental snapshot inode file. Blocks 5 and 6 are unchanged in both files, and thus, are quickly scanned over without further processing of any inodes or other information. Hence, scanned blocks at the same index in both snapshots can be efficiently bypassed, reducing the scan time.

[0053] Block pairs (e.g. blocks 7 and 8) that have been identified as changed are forwarded (as they are detected by the scan/scanner 820) to the rest of the source process, which includes an inode picker process 830. The inode picker identifies specific inodes (based upon qtree ID) from the forwarded blocks that are part of the selected qtree being mirrored. In this example the qtree

ID Q2 is selected, and inodes containing this value in their file metadata are "picked" for further processing. Other inodes not part of the selected qtree(s) (e.g. inodes with qtree IDs Q1 and Q3) are discarded or otherwise ignored by the picker process 830. Note that a multiplicity of qtree IDs can be selected, causing the picker to draw out a group of inodes—each having one of the selected qtree associations.

[0054] The appropriately "picked" inodes from changed blocks are then formed into a running list or queue 840 of changed inodes 842. These inodes are denoted by a discrete inode number as shown. Each inode in the queue 840 is handed off to an inode handler or worker 850, 852 and 854 as a worker becomes available. Fig. 8A is a table 835 detailing the basic set of rules the inode picker process 830 uses to determine whether to send a given inode to the queue for the workers to process.

[0055] The mode picker process 830 queries whether either (1) the base snapshot's version of the subject inode (a given inode number) is allocated *and* in a selected qtree (box 860) or (2) the incremental snapshot's version of the inode is allocated *and* in a selected qtree (box 862). If neither the base nor incremental version are allocated and in the selected qtree then both inodes are ignored (box 864) and the next pair of inode versions are queried.

[0056] If the base inode is not in allocated or not in the selected qtree, but the incremental inode is allocated and in the selected qtree, then this implies an incremental file has been added, and the appropriate inode change is sent to the workers (box 866). Similarly, if the base inode is allocated and in the selected qtree, but the incremental inode is not allocated or not in the selected qtree, then this indicates a base file has been deleted and this is sent on to the destination via the data stream format (as described below) (box 868).

[0057] Finally, if a base inode and incremental inode are both allocated and in the selected qtree, then the process queries whether the base and incremental inodes represent the same file (box 870). If they represent the same file, then the file or its metadata (permissions, owner, permissions, etc) *may* have changed. This is denoted by different generation numbers on different versions of the inode number being examined by the picker process. In this case, a *modified* file is sent and the inode workers compare versions to determine exact changes as described further below (box 872). If the base and incremental are not the exact same file, then this implies a deletion of the base file and addition of an incremental file (box 874). The addition of the incremental file is noted as such by the picker in the worker queue.

[0058] Fig. 8B is a more detailed view of the information contained in exemplary changed blocks (block 10) in the base snapshot 810 and (block 12) in the incremental snapshot 812, respectively. Inode 2800 is unallocated in the base inode file and allocated in the incremental inode file. This implies that the file has been add-

ed to the file system. The inode picker process also notes that this inode is in the proper qtree Q2 (in this example). This inode is sent to the changed inode queue for processing, with a note that the whole file is new.

[0059] Inode 2801 is allocated in both inode files. It is in the proper qtree Q2, and the two versions of this inode share the same generation number. This means that the inode represents the same file in the base and the incremental snapshots. It is unknown at this point whether the file data itself has changed, so the inode picker sends the pair to the changed inode queue, and a worker determines what data has changed. Inode 2802 is allocated in the base inode file, but not allocated in the incremental inode file. The base version of the inode was in the proper qtree Q2. This means this inode has been deleted. The inode picker sends this information down to the workers as well. Finally, inode 2803 is allocated in the base inode file, and *reallocated* in the incremental inode file. The inode picker 830 can determine this because the generation number has changed between the two versions (from #1 - #2). The new file which this inode represents has been added to the qtree, so like inode 2800, this is sent to the changed inode queue for processing, with a note that the whole file is new.

[0060] A predetermined number of workers operate on the queue 840 at a given time. In the illustrative embodiment, the workers function in parallel on a group of modes in the queue. That is, the workers process inodes to completion in no particular order once taken from the queue and are free process further inodes from the queue as soon as they are available. Other processes, such as the scan 820 and picker 830 are also interleaved within the overall order.

[0061] The function of the worker is to determine changes between each snapshot's versions of the files and directories. As described above, the source snapshot mirror application is adapted to analyze two versions of inodes in the two snapshots and compares the pointers in the inodes. If the two versions of the pointers point to the same block, we know that that block hasn't changed. By extension, if the pointer to an indirect block has not changed, then that indirect block has not changed data, so none of its pointers can have changed, and, thus, *none of the data blocks underneath it* in the tree have changed. This means that, in a very large file, which is mostly unchanged between two snapshots, the process can skip over/overlook VBN "pointers" to each data block in the tree to query whether the VBNs of the data blocks have changed.

[0062] The operation of a worker 850 is shown by way of example in Fig. 9. Once a changed inode pair are received by the worker 850, each inode (base and incremental, respectively) 910 and 912 is scanned to determine whether the file offset between respective blocks is a match. In this example, blocks 6 and 7 do not match. The scan then continues down the "tree" of blocks 6 and 7, respectively, arriving at underlying indirect blocks 8/9 (920) and 8/10 (922). Again the file offset comparison

indicates that blocks 8 both arrive at a common block 930 (and thus have not changed). Conversely, blocks 9 and 10 do not match due to offset differences and point to changed blocks 940 and 942. The changed block 942 and the metadata above can be singled out for transmission to the replicated snapshot on the destination (described below; see also Fig. 8). The tree, in an illustrative embodiment extends four levels in depth, but this procedure may be applied to any number of levels. In addition, the tree may in fact contain several changed branches, requiring the worker (in fact, the above-described scanner 820 process) to traverse each of the branches in a recursive manner until all changes are identified. Each inode worker, thus provides the changes to the network for transmission in a manner also described below. In particular, new blocks and information about old, deleted blocks are sent to the destination. Likewise, information about modified blocks is sent.

[0063] Notably, because nearly every data structure in this example is a file, the above-described process can be applied not only to file data, but also to directories, access control lists (ACLs) and the inode file itself.

[0064] It should be again noted, that the source procedure can be applied to any level of granularity of file system organization, including an entire volume inode file. By using the inherent qtree organization a quick and effective way to replicate a known subset of the volume is provided.

2. Communication Between Source and Destination

[0065] With further reference to Fig. 10, the transmission of changes from the source snapshot to the replicated destination snapshot is described in an overview 1000. As already described, the old and new snapshots present the inode picker 830 with changed inodes corresponding to the qtree or other selected sub-organization of the subject volume. The changed inodes are placed in the queue 840, and then their respective trees are walked for changes by a set of inode workers 850, 852 and 854. The inode workers each send messages 1002, 1004 and 1006 containing the change information to a source pipeline 1010. Note that this pipeline is only an example of a way to implement a mechanism for packaging file system data into a data stream and sending that stream to a network layer. The messages are routed first to a receiver 1012 that collects the messages and sends them on to an assembler 1014 as a group comprising the snapshot change information to be transmitted over the network 318. Again, the "network" as described herein should be taken broadly to include anything that facilitates transmission of volume sub-organization (e.g. qtree) change data from a source sub-organization to a destination sub-organization, even where source and destination are on the same file server, volume or, indeed (in the case of rollback as described in the above-mentioned U.S. Patent Application entitled SYSTEM AND METHOD FOR REMOTE

ASYNCHRONOUS MIRRORING USING SNAPSHOTS) are the same sub-organization at different points in time. An example of a "network" used as a path back to the same volume is a loopback. The assembler 1014 generates a specialized format 1020 for transmitting the data stream of information over the network 318 that is predictable and understood by the destination. The networker 1016 takes the assembled data stream and forwards it to a networking layer. This format is typically encapsulated within a reliable networking protocol such as TCP/IP. Encapsulation can be performed by the networking layer, which constructs, for example, TCP/IP packets of the formatted replication data stream

[0066] The format 1020 is described further below. In general, its use is predicated upon having a structure that supports multiple protocol attributes (e.g. Unix permissions, NT access control lists (ACLs), multiple file names, NT streams, file type, file-create/modify time, etc.). The format should also identify the data in the stream (i.e. the offset location in a file of specific data or whether files have "holes" in the file offset that should remain free). The names of files should also be relayed by the format. More generally, the format should also be independent of the underlying network protocol or device (in the case of a tape or local disk/non-volatile storage) protocol and file system—that is, the information is system "agnostic," and not bound to a particular operating system software, thereby allowing source and destination systems of different vendors to share the information. The format should, thus, be completely self-describing requiring no information outside the data stream. In this manner a source file directory of a first type can be readily translated into destination file directory of a different type. It should also allow extensibility, in that newer improvements to the source or destination operating system should not affect the compatibility of older versions. In particular, a data set (e.g. a new header) that is not recognized by the operating system should be ignored or dealt with in a predictable manner without triggering a system crash or other unwanted system failure (i.e. the stream is backwards compatible). This format should also enable transmission of a description of the whole file system, or a description of only changed blocks/information within any file or directory. In addition, the format should generally minimize network and processor overhead.

[0067] As changed information is forwarded over the network, it is received at the destination pipeline piece 1030. This pipeline also includes a networker 1032 to read out TCP/IP packets from the network into the snapshot replication data stream format 1020 encapsulated in TCP/IP. A data reader and header stripper 1034 recognizes and responds to the incoming format 1020 by acting upon information contained in various format headers (described below). A file writer 1036 is responsible for placing file data derived from the format into appropriate locations on the destination file system.

[0068] The destination pipeline 1030 forwards data

and directory information to the main destination snapshot mirror process 1040, which is described in detail below. The destination snapshot mirror process 1040 consists of a directory stage 1042, which builds the new replicated file system directory hierarchy on the destination side based upon the received snapshot changes. To briefly summarize, the directory stage creates, removes and moves files based upon the received formatted information. A map of inodes from the destination to the source is generated and updated. In this manner, inode numbers on the source file system are associated with corresponding (but typically different) inode numbers on the destination file system. Notably, a temporary or "purgatory" directory 1050 (described in further detail below) is established to retain any modified or deleted directory entries 1052 until these entries are reused by or removed from the replicated snapshot at the appropriate directory rebuilding stage within the directory stage. In addition, a file stage 1044 of the destination mirror process populates the established files in the directory stage with data based upon information stripped from associated format headers.

[0069] The format into which source snapshot changes are organized is shown schematically in Figs. 11 and 12. In the illustrative embodiment, the format is organized around 4 KB blocks. The header size and arrangement can be widely varied in alternate embodiments, however. There are 4 KB headers (1100 in Fig. 11) that are identified by certain "header types." Basic data stream headers ("data") are provided for at most every 2 megabytes (2 MB) of *changed* data. With reference to Fig. 11, the 4 KB standalone header includes three parts, a 1 KB generic part 1102, a 2 KB non-generic part 1104, and an 1 KB expansion part. The expansion part is not used, but is available for later versions.

[0070] The generic part 1102 contains an identifier of header type 1110. Standalone header types (i.e. headers not followed by associated data) can indicate a start of the data stream; an end of part one of the data stream; an end of the data stream; a list of deleted files encapsulated in the header; or the relationship of any NT *streamdirs*. Later versions of Windows NT allow for multiple NT "streams" related to particular filenames. Also in the generic part 1102 is a checksum 1112 that ensures the header is not corrupted. In addition other data such as a "checkpoint" 1114 used by the source and destination to track the progress of replication is provided. By providing a list of header types, the destination can more easily operate in a backwards-compatible mode—that is, a header type that is not recognized by the destination (provided from a newer version of the source) can be more easily ignored, while recognized headers within the limits of the destination version are processed as usual.

[0071] The kind of data in the non-generic part 1104 of the header 1100 depends on the header type. It could include information relating to file offsets (1120) in the case of the basic header, used for follow-on data trans-

mission, deleted files (in a standalone header listing of such files that are no longer in use on the source or whose generation number has changed) (1122), or other header-specific information (1124 to be described below). Again, the various standalone headers are interposed within the data stream format at an appropriate location. Each header is arranged to either reference an included data set (such as deleted files) or follow-on information (such as file data).

[0072] Fig. 12 describes the format 1020 of the illustrative replication data stream in further detail. The format of the replicated data stream is headed by a standalone data stream header 1202 of the type "start of data stream." This header contains data in the non-generic part 1104 generated by the source describing the attributes of the data stream.

[0073] Next a series of headers and follow-on data in the format 1020 define various "part 1" information (1204). Significantly, each directory data set being transmitted is preceded by a basic header with no non-generic data. Only directories that have been modified are transmitted, and they need not arrive in a particular order. Note also that the data from any particular directory need not be contiguous. Each directory entry is loaded into a 4 KB block. Any overflow is loaded into a new 4 KB block. Each directory entry is a header followed by one or more names. The entry describes an inode and the directory names to follow. NT stream directories are also transmitted.

[0074] The part 1 format information 1204 also provides ACL information for every file that has an associated ACL. By transmitting the ACLs before their associated file data, the destination can set ACLs before file data is written. ACLs are transmitted in a "regular" file format. Deleted file information (described above) is sent with such information included in the non-generic part 1104 of one or more standalone headers (if any). By sending this information in advance, the directory tree builder can differentiate between moves and deletes.

[0075] The part 1 format information 1204 also carries NT stream directory (streamdir) relationship information. One or more standalone headers (if any) notifies the destination file server of every changed file or directory that implicates NT streams, regardless of whether the streams have changed. This information is included in the non-generic part 1104 of the header 1100 (Fig. 11).

[0076] Finally, the part 1 format information 1204 includes special files for every change in a symlink, named pipe, socket, block device, or character device in the replicated data stream. These files are sent first, because they are needed to assist the destination in building the infrastructure for creation of the replicated file system before it is populated with file data. Special files are, like ACLs, transmitted in the format of regular files.

[0077] Once various part 1 information 1204 is transmitted, the format calls for an "end of part 1 of the data

stream" header 1206. This is a basic header having no data in the non-generic part 1104. This header tells the destination that part 1 is complete and to now expect file data.

5 [0078] After the part 1 information, the format presents the file and stream data 1208. A basic header 1210 for every 2 MB or less of *changed* data in a file is provided, followed by the file data 1212 itself. The files comprising the data need not be written in a particular order, nor must the data be contiguous. In addition, referring to the header in Fig. 11, the basic header includes a block numbers data structure 1130, associated with the non-generic part 1104 works in conjunction with the "holes array" 1132 within (in this example) the generic part 1102. The holes array denotes empty space. This structure, in essence, provides the mapping from the holes array to corresponding blocks in the file. This structure instructs the destination where to write data blocks or holes.

10 [0079] In general files (1212) are written in 4 KB chunks with basic headers at every 512 chunks (2 MB), at most. Likewise, streams (also 1212) are transmitted like regular files in 4 KB chunks with at most 2 MB between headers.

15 [0080] Finally, the end of the replicated data stream format 1020 is marked by a footer 1220 consisting of standalone header of the type "end of data stream." This header has no specific data in its non-generic part 1104 (Fig. 11).

3. Destination

[0081] When the remote destination (e.g. a remote file server, remote volume, remote qtree or the same qtree) receives the formatted data stream from the source file server via the network, it creates a new qtree or modifies an existing mirrored qtree (or another appropriate organizational structure) and fills it with data. Fig. 13 shows the destination snapshot mirror process 1040 in greater detail. As discussed briefly above, the process consists of two main parts, a directory stage 1042 and a data or file stage 1044.

[0082] The directory stage 1042 is invoked first during a transmission the data stream from the source. It consists of several distinct parts. These parts are designed to handle all part 1 format (non-file) data. In an illustrative embodiment the data of part 1 is read into the destination, stored as files locally, and then processed from local storage. However, the data may alternatively be processed as it arrives in realtime.

[0083] More particularly, the first part of the directory stage 1042 involves the processing of deleted file headers (1310). Entries in the inode map (described further below) are erased with respect to deleted files, thereby severing a relation between mapped inodes on the replicated destination snapshot and the source snapshot.

[0084] Next the directory stage undertakes a tree cleaning process (1312). This step removes all directory

entries from the replicated snapshot directory 1330 that have been changed on the source snapshot. The data stream format (1020) indicates whether a directory entry has been added or removed. In fact, directory entries from the base version of the directory and directory entries from the incremental version of the directory are both present in the format. The destination snapshot mirror application converts the formatted data stream into a destination directory format in which each entry that includes an inode number, a list of relative names (e.g. various multi-protocol names) and a "create" or "delete" value. In general each file also has associated therewith a generation number. The inode number and the generation number together form a *tuple* used to directly access a file within the file system (on both the source and the destination). The source sends this tuple information to the destination within the format and the appropriate tuple is stored on the destination system. Generation numbers that are out of date with respect to existing destination files indicate that the file has been deleted on the source. The use of generation numbers is described further below.

[0085] The destination processes base directory entries as removals and incremental directory entries as additions. A file which has been moved or renamed is processed as a delete (from the old directory or from the old name), then as an add (to the new directory or with a new name). Any directory entries 1052 that are deleted, or otherwise modified, are moved temporarily to the temporary or "purgatory" directory, and are not accessible in this location by users. The purgatory directory allows modified entries to be, in essence, "moved to the side" rather than completely removed as the active file system's directory tree is worked on. The purgatory directory entries, themselves point to data, and thus prevent the data from becoming deleted or losing a link to a directory altogether.

[0086] On a base transfer of a qtree to the destination, the directory stage tree building process is implemented as a breadth-first traversal of all the files and directories in the data stream, starting with the root of the qtree. The directory stage then undertakes the tree building process, which builds up all the directories with stub entries for the files. However, the depicted incremental directory stage (1042), as typically described herein, differs from a base transfer in that the tree building process (1314) begins with a directory queue that includes *all* modified directories currently existing on both the source and the destination (i.e. the modified directories that existed prior to the transfer). The incremental directory stage tree building process then processes the remainder of the directories according to the above-referenced breadth-first approach.

[0087] For efficiency, the source side depends upon inode numbers and directory blocks rather than pathnames. In general, a file in the replicated directory tree (a qtree in this example) on the destination cannot expect to receive the same inode number as the corre-

sponding file has used on the source (although it is possible). As such, an *inode map* is established in the destination. This map 1400, shown generally in Fig. 14, enables the source to relate each file on the source to the destination. The mapping is based generally upon file offsets. For example a received source block having "offset 20KB in inode 877" maps to the block at offset 20 KB in replicated destination inode 9912. The block can then be written to the appropriate offset in the destination file.

[0088] More specifically, each entry in the inode map 1400 contains an entry for each inode on the source snapshot. Each inode entry 1402 in the map is indexed and accessed via the source inode number (1404). These source inodes are listed in the map in a sequential and monotonically ascending order, notwithstanding the order of the mapped destination inodes. Under each source inode number (1404), the map includes: the source generation number (1406) to verify that the mapped inode matches the current file on the source; the destination inode number (1408); and destination generation number (1410). As noted above, the inode number and generation number together comprise a *tuple* needed to directly access an associated file in the corresponding file system.

[0089] By maintaining the source generation number, the destination can determine if a file has been modified or deleted on the source (and its source associated inode reallocated), as the source generation number is incremented upwardly with respect to the stored destination. When the source notifies the destination that an inode has been modified, it sends the tuple to the destination. This tuple uniquely identifies the inode on the source system. Each time the source indicates that an entirely new file or directory has to be created (e.g. "create") the destination file system creates that file. When the file is created, the destination registers data as a new entry in its inode map 1400. Each time the source indicates that an existing file or directory needs to be deleted, the destination obliterates that file, and then clears the entry in the inode map. Notably, when a file is modified, the source only sends the tuple and the data to be applied. The destination loads the source inode's entry from the inode map. If the source generation number matches, then it knows that the file already exists on the destination and needs to be modified. The destination uses the tuple recorded in the inode map to load the destination inode. Finally, it can apply the file modifications by using the inode.

[0090] As part of the tree building process reused entries are "moved" back from the purgatory directory to the replicated snapshot directory 1330. Traditionally, a move of a file requires knowledge of the name of the moved file and the name of the file it is being moved to. The original name of the moved file may not be easily available in the purgatory directory. In addition, a full move would require two directories (purgatory and replicated snapshot) to be modified implicating additional

overhead.

[0091] However, in the illustrative embodiment, if the source inodes received at the destination refer to inodes in the inode map 1400, then the directory stage creates (on the current built-up snapshot directory 1330) a file entry having the desired file name. This name can be exactly the name derived from the source. A hard link 1332 (i.e. a Unix-based link enables multiple names to be assigned to a discrete file) is created between that file on the snapshot directory 1330 and the entry in the purgatory directory. By so linking the entry, it is now pointed to by both the purgatory directory and the file on the snapshot directory itself. When the purgatory directory root is eventually deleted (thereby killing off purgatory) at the end of the data stream transfer, the hard link will remain to the entry, ensuring that the specific entry in the purgatory directory will not be deleted or recycled (given that the entry's link count is still greater than zero) and a path to the data from the file on the new directory is maintained. Every purgatory entry that eventually becomes associated with a file in the newly built tree will be similarly hard linked, and thereby survive deletion of the purgatory directory. Conversely, purgatory entries that are not relinked will not survive, and are effectively deleted permanently when purgatory is deleted.

[0092] It should now be clear that the use of mapping and generation number tuples avoids the expensive (from a processing standpoint) use of conventional full file pathnames (or relative pathnames) in the data stream from the source. Files that are modified on the source can be updated on the destination without loading a directory on either the source or destination. This limits the information needed from the source and the amount of processing required. In addition, the source need not maintain a log of directory operations. Likewise, since the destination need not maintain a central repository of the current file system state, multiple sub-directories can be operated upon concurrently. Finally, neither the source, nor the destination must explicitly track deleted files as such deleted files are automatically removed. Rather, the source only sends its list of deleted files and the destination uses this list to conform the inode map. As such, there is no need to selectively traverse a tree more than once to delete files, and at the conclusion of the transfer, simply eliminating the purgatory directory is the only specific file cleaning step.

[0093] The directory stage 1042 sets up any ACLs on directories as the directories are processed during tree building (substep 1316). As described above, the ACL and NT stream relationships to files are contained in appropriate standalone headers. ACLs are then set on files during the below-described file stage. NT streams are created on files as the files are, themselves, created. Since an NT stream is, in fact, a directory, the entries for it are processed as part of the directory phase.

[0094] The new directory tree may contain files with no data or old data. When the "end of part 1" format header is read, the destination mirror process 1040 en-

ters the file stage 1044 in which snapshot data files 1340 referenced by the directory tree are populated with data (e.g. change data). Fig. 15 shows a simplified procedure 1500 for writing file data 1502 received from the source.

In general, each (up to) 2 MB of data in 4 KB blocks arrives with corresponding source inode numbers. The inode map 1400 is consulted for corresponding entries 1402. Appropriate offsets 1504 are derived for the data, and it is written into predetermined empty destination snapshot data files 1340.

[0095] At the end of both the directory stage 1042 and data stage 1044, when all directory and file data have been processed, and the data stream transfer from the source is complete, the new replicated snapshot is exposed atomically to the user. At this time the contents of the purgatory directory 1050 (which includes any entries that have not be "moved" back into the rebuilt tree) is deleted.

[0096] It should be noted that the initial creation (the "level zero" transfer) of the replicated snapshot on the destination follows the general procedures discussed above. The difference between a level zero transfer and a regular update is that there is no base snapshot; so the comparisons always process information in the incremental snapshot as additions and creates rather than modifications. The destination mirror application starts tree building by processing any directories already known to it. The initial directory established in the destination is simply the root directory of the replicated snapshot (the qtree root). A destination root exists on the inode map. The source eventually transmits a root (other files received may be buffered until the root arrives), and the root is mapped to the existing destination root. Files referenced in the root are then mapped in turn in a "create" process as they are received and read by the destination. Eventually, the entire directory is created, and then the data files are populated. After this, a replica file system is complete.

40 E. Rollback

[0097] As described above, a source and destination can be the same qtree, typically at different points in time. In this case, it is contemplated that an incremental change to a snapshot can be undone by applying a "rollback" procedure. In essence, the base and incremental snapshot update process described above with reference to Fig. 8 is performed in reverse so as to recover from a disaster, and return the active file system to the state of a given snapshot.

[0098] Reference is made to Fig. 16, which describes a generalized rollback procedure 1600 according to an illustrative embodiment. As a matter of ongoing operation, in step 1605, a "first" snapshot is created. This first snapshot may be an exported snapshot of the replicated snapshot on the destination. In the interim, the subject destination active file system (replicated snapshot) is modified by an incremental update from the source (step

1610).

[0099] In response to an exigency, such as a panic, crash, failure of the update to complete or a user-initiated command, a rollback initiation occurs (step 1615). This is a condition in which the next incremental update of the replicated snapshot will not occur properly, or otherwise does not reflect an accurate picture of the data.

[0100] In response to rollback initiation, further modification/update to the replicated snapshot is halted or frozen (step 1620). This avoids further modifications that may cause the active file system to diverge from the state to be reflected in a second snapshot that will be created from the active file system in the next step (step 1625 below) immediately after the halt. Modification to the active file system is halted using a variety of techniques such as applying read only status to the file system or denying all access. In one embodiment, access to the active file system is redirected to an exported snapshot by introducing a level of indirection to the inode lookup of the active file system, as set forth in the above-mentioned U.S. Patent Application entitled SYSTEM AND METHOD FOR REDIRECTING ACCESS TO A REMOTE MIRRORED SNAPSHOT.

[0101] After the halt, a "second" exported snapshot of the modified active file system in its most current state is now created (step 1625).

[0102] Next, in step 1630, the incremental changes are computed between the second and the first snapshots. This occurs in accordance with the procedure described above with reference to Figs. 8 and 9, but using the second snapshot as the base and the first snapshot as the incremental. The computed incremental changes are then applied to the active file system (now frozen in its present state) in step 1635. The changes are applied so that the active file system is eventually "rolled back" to the state contained in the first snapshot (step 1640). This is the active file system state existing before the exigency that necessitated the rollback.

[0103] In certain situations, the halt or freeze on further modification of the active file system according to step 1625 is released, allowing the active file system to again be accessed for modification or user intervention (step 1645). However, in the case of certain processes, such as rollback (described below), a rolled back qtree is maintained under control for further modifications by the replication process.

[0104] One noted advantage to the rollback according to this embodiment is that it enables the undoing of set of changes to a replicated data set without the need to maintain separate logs or consuming significant system resources. Further the direction of rollback—past-to-present or present-to-past—is largely irrelevant. Furthermore, use of the purgatory directory, and not deleting files, enables the rollback to not affect existing NFS clients. Each NFS client accesses files by means of file handles, containing the inode number and generation of the file. If a system deletes and recreates a file, the file will have a different inode/generation tuple. As such,

the NFS client will not be able to access the file without reloading it (it will see a message about a stale file handle). The purgatory directory, however, allows a delay in unlinking files until the end of the transfer. As such, a rollback as described above can resurrect files that have just been moved into purgatory, without the NFS clients taking notice.

F. Inode Map Flip

[0105] Where a destination replicated snapshot may be needed at the source to, for example, rebuild the source qtree snapshot, (in other words, the role of the source and destination snapshot are reversed) the use of generalized rollback requires that the inode map be properly related between source and destination. This is because the source inodes do not match the destination inodes in their respective trees. For the same reason an inode map is used to construct the destination tree, the source must exploit a mapping to determine the nature of any inodes returned from the destination during the rollback. However, the inode map residing on the destination does not efficiently index the information in a form convenient for use by the source. Rather, the source would need to hunt randomly through the order presented in the map to obtain appropriate values.

[0106] One way to provide a source-centric inode map is to perform a "flip" of map entries. Fig. 17 details a procedure 1700 for performing the flip. The flip operation is initiated (step 1705) as part of a rollback initiated as part of a disaster recovery procedure or for other reasons (automatically or under user direction). Next, the destination and source negotiate to transfer the inode map file to the source from the destination. The negotiation can be accomplished using known data transfer methodologies and include appropriate error correction and acknowledgements (step 1710). The inode is thereby transferred to the source from the destination and is stored.

[0107] Next the source (which after the negotiation becomes the new destination), creates an empty inode map file with one entry for each inode in the source qtree (step 1715). The new destination then initializes a counter with (in this example) $N=1$ (step 1720). N is the variable representing the inode count on the new destination qtree.

[0108] In step 1725, the new destination looks up the N th inode from the entries associated with the old destination in the stored inode map file (i.e. the map from the old destination/new source). Next, the new destination determines if such an entry exists (decision step 1730). If no entry exists, then a zero entry is created in the new inode map file, representing that the N th inode of the new source (old destination) is not allocated. However, if there exists an N th inode of the new source/old destination, then the decision step 1730 branches to step 1740, and creates a new entry in the new inode map file (created in step 1715). The new entry maps the

new source (old destination) Nth inode to the proper new destination (old source) inode. Note, in an alternate embodiment, the new inode map is provided with a full field of zero entries before the mapping begins, and the creation of a "zero entry," in this case should be taken broadly to include leaving a preexisting zero entry in place in the inode map.

[0109] The procedure 1700 then checks if N equals the number of inodes in the old destination file system (decision step 1745). If so, the new inode map file is complete and the procedure quits (step 1750). Conversely, if additional modes are still to-be-mapped, then the counter is incremented by one ($N=N+1$ in step 1755). Similarly, if a zero entry is made into the new inode map, then the procedure 1700 also branches to decision step 1745 to either increment the counter (step 1755) or quit (step 1750). Where the counter is incremented in step 1755, the procedure branches back to step 1725 wherein the incremented Nth inode is looked up.

[0110] By way of example, Fig. 18 shows an illustrative old destination inode map file 1800 including three exemplary entries 1802, 1804 and 1806, sequentially. The fields 1404, 1406 (source and destination inode numbers), 1408, 1410 (source and destination generation numbers) are described above with reference to Fig. 14. Entry 1802 shows that (old) source inode 72 maps to (old) destination inode 605. Likewise entry 1804 maps source inode 83 to destination inode 328, and entry 1806 maps source inode 190 to destination inode 150.

[0111] Fig 19 shows an exemplary new inode map file 1900 generated from the old inode map file 1800 of Fig. 18 in accordance with the flip procedure 1700. The new map includes fields for the new source (old destination) inode 1902, new destination (old source) inode 1904, new source (old destination) generation number 1906 and new destination (old source) generation number 1908. As a result of the flip, the entry 1910 for new source inode 150 is presented in appropriate index order and is paired with new destination inode 190 (and associated generation numbers). The entry 1912 for new source inode 328 is next (after a series of consecutive, intervening entries 1914 for new source inodes 151-372) and maps new destination inode 83. Likewise the entry 1916 for new source inode 605 maps new destination inode 72, after intervening entries 1918 for new source inodes 329-604. The intervening source inodes may contain mappings to other new existing destination inodes, or they may have a zero value as shown in entry 1930 for new source inode 606 (as provided by step 1735 of the procedure 1700 where no new destination inode was detected on the stored old source inode map (1800)).

G. Inode Map Association

[0112] It is further contemplated that, two replica/mir-

ror snapshots of the same source can establish a mirror relationship with one another. These two snapshots may be representative of two different points in time with respect to the original source. Fig. 20 shows a generalized environment 2000 in which an original source 2001 has generated two replica/mirror snapshots Destination Snapshot A (2002) and Destination Snapshot B (2004). Each Destination Snapshot A and B (2002 and 2004) has an associated mode map A and B (2012 and 2014, respectively), used to map the inodes of transferred data stream from the original source 2001.

[0113] In the illustrated example, the Destination Snapshot A (2002) is now prepared to transfer changes so as to establish a mirror in Destination Snapshot B (2004). However, the reverse is also contemplated, i.e. Destination Snapshot B establishing a Mirror in Destination Snapshot A. Thus, Destination Snapshot A (2002) becomes the new "source" in the transfer with Destination Snapshot B (2004) acting as the desired destination system for replication data from Destination Snapshot A. As in the above-described flip embodiment, the new source 2002 transfers its inode map A 2012 to the destination system 2004. The destination system 2004 then determines the relationship between the two system's inodes. In this case, both the new source and the new destination system have their own inode maps A and B (2012 and 2014), indexed off the old source 2001, and referencing the inodes in their respective trees. Given the existence of two respective inode maps, an "associative" process 2016 walks the inode maps concurrently, inode-by-inode. For each inode from the original source 2001, the process extracts the "destination inode/generation number" from each of the inode maps A and B. It then treats the new source as the appropriate map index for the new associated inode map 2018. In the associated map, it stores the new source generation number for the new source index inode number, with each index entry also associated with/ mapped to the new destination inode/generation number extracted from the inode map B (2014). The new map is used by the new destination 2004 in accordance with the principles described above to build trees in the directory based upon changes in the new source with respect to various points in time.

[0114] By way of example, an hypothetical old source OS inode number 55 (OS 55) is mapped to old destination snapshot A in its map A to old destination A inode 87 (A 87) and OS 55 is mapped to old destination B inode 99 (B 99) in map B. To make B the new destination and A the new source, an associative map is constructed with the process extracting A 87 and B 99 for the respective maps based upon the common index OS 55. The associated map contains the new source/new destination entry 87/99. It also includes the associated generation numbers with these values from the old maps A and B. Note that, while the procedure is applied to two old destination systems, it is contemplated that more than two destination systems can be associated in var-

ious ways in accordance with the techniques described herein.

[0115] The foregoing has been a detail description of illustrative embodiments of the invention. Various modifications and additions can be made without departing from the scope of the invention. For example, the number of interconnected source and/or destination servers depicted can be varied. In fact, the source and destination servers can be the same machine. It is expressly contemplated that a plurality of sources can transfer data to a destination and *vice versa*. Likewise, the internal architecture of the servers or their respective storage arrays, as well as their network connectivity and protocols, are all highly variable. The operating systems used on various source and destination servers can differ. In addition, it is expressly contemplated that any of the operations and procedures described herein can be implemented using hardware, software comprising a computer-readable medium having program instructions executing on a computer, or a combination of hardware and software. The computer readable medium can be any suitable carrier medium such as a signal e.g. an electrical, optical, microwave, magnetic, acoustic or electromagnetic signal, or a storage medium such as a floppy disk, hard disk, CD ROM or programmable memory device.

[0116] In one aspect the present invention provides a system for generating a group of changes in a source file system snapshot pair for transmission to a replicated destination file system comprising: a scanner that searches a respective logical block file index of each of a first snapshot and a second snapshot of the source file system according to a predetermined file index and that retrieves only blocks having volume block numbers at corresponding file index locations on each of the first snapshot and the second snapshot that have different volume block numbers; wherein the scanner is adapted to walk down a hierarchy of pointers associated with each respective logical block file index and retrieve only pointed-to blocks with changed volume block numbers relative to pointed-to blocks at corresponding file offsets while bypassing blocks having unchanged volume block numbers therebetween.

[0117] In one embodiment an inode picker picks out inodes in the retrieved blocks having a predetermined association and that indicate changes between, respectively a first version and a second version of the inodes in the first snapshot and the second snapshot.

[0118] In one embodiment the predetermined association comprises a sub-organization of a volume file system.

[0119] In one embodiment the sub-organization comprises a qtree within the volume file system.

[0120] In one embodiment an inode worker operates on queued blocks retrieved by the inode picker and that applies the scanner to derive changes between blocks in a hierarchy of the first version of each of the picked out inodes with respect to a hierarchy of the second ver-

sion of each of the picked out inodes.

[0121] In one embodiment the inode worker is adapted to collect the changes and forward the changes to a pipeline that packages the changes in a data stream for transmission to the destination file system.

[0122] In one embodiment an extensible and file system-independent format transmits the changes between the source file system and the destination file system.

[0123] In one embodiment the format comprises a plurality of standalone headers including directory information, including deleted files, and basic data file headers, following the standalone headers, each of the basic data headers being followed by a stream of file data.

[0124] In one embodiment the destination file system includes an inode map for mapping inode numbers and associated file generation numbers provided by the pipeline to corresponding inode numbers and associated file generation numbers recognized by the destination file system.

[0125] In one embodiment at the destination file system, a temporary directory, is adapted to receive thereinto any directory entries from the destination file system for files indicated by the changes as modified and deleted files with respect to corresponding entries for the files on the inode map, the temporary directory being deleted at a conclusion an update of the destination file system with the changes, wherein any modified files and deleted files that remain associated with the destination file system remain linked to the destination file system.

[0126] In one embodiment the inode worker is part of a plurality of like inode workers that each operate on queued blocks retrieved by the inode picker and that apply the scanner to derive changes between blocks in a hierarchy of the first version of each of the picked out inodes with respect to a hierarchy of the second version of each of the picked out inodes.

[0127] Another aspect of the invention provides a method for generating a group of changes in a source file system snapshot pair for transmission to a replicated destination file system comprising: searching, with a scanner, a respective logical block file index of each of a first snapshot and a second snapshot of the source file system according to a predetermined file index and retrieving only blocks having volume block numbers at corresponding file index locations on each of the first snapshot and the second snapshot that have different volume block numbers; and with the scanner, walking down a hierarchy of pointers associated with each respective logical block file index and retrieve only pointed-to blocks with changed volume block numbers relative to pointed-to blocks at corresponding file offsets while bypassing blocks having unchanged volume block numbers therebetween.

[0128] In one embodiment the method includes picking out inodes, with an inode picker, in the retrieved blocks having a predetermined association and indicating changes between, respectively a first version and a

second version of the inodes in the first snapshot and the second snapshot.

[0129] In one embodiment the predetermined association comprises a sub-organization of a volume file system.

[0130] In one embodiment the sub-organization comprises a qtree within the volume file system.

[0131] In one embodiment the method further comprises operating on queued blocks retrieved by the inode picker with an inode worker that applies the scanner to derive changes between blocks in a hierarchy of the first version of each of the picked out inodes with respect to a hierarchy of the second version of each of the picked out inodes.

[0132] A further aspect of the invention provides a system for deriving changes in a base and incremental snapshot of a source file system and transferring the changes to update a replicated destination file system comprising: a source scanning process that compares volume block numbers of inodes for the base and the incremental snapshots along respective logical file indexes therefor, and that returns changes in pointed-to blocks with respect to a predetermined association of the inodes; a format that packages the changes in a form that provides directory information and file data information under separate headers; and a destination directory process that builds a directory of the replicated destination file system by mapping, with an inode map, source inode numbers and associated file generation numbers to destination file system inode numbers and associated generation numbers.

[0133] In one embodiment a temporary directory of the destination file system receives any modified and deleted files from the source with respect to existing files in the inode map and that is adapted to be deleted after an update of the destination file system is complete wherein any remaining files remain linked to the directory of the replicated destination file system.

[0134] In one embodiment the predetermined association comprises a sub-organization of a volume.

[0135] In one embodiment the sub-organization comprises a qtree.

Claims

1. A system for identifying changes in a logical group of data blocks on a source and updating a replica of the logical group of data blocks on a destination, the system comprising:

a scanner operative on a first set of block identifiers and a second set of block identifiers of a first snapshot and second snapshot respectively, the first snapshot and the second snapshot each respectively corresponding to a different image of the logical group of data blocks with the first snapshot corresponding to the replica,

the scanner being adapted to search the second set of block identifiers for block identifiers that differ from corresponding block identifiers of the first set of block identifiers and thereby indicate changed data blocks that are not reflected in the replica;

transmitting the changed blocks to the destination without transmitting all the blocks of the logical group; and

updating the replica on the destination with the changed blocks.

2. The system of claim 1, wherein the scanner is adapted to walk down a hierarchy of pointers associated with respective logical block indexes for the first and second snapshots and identify pointed-to blocks with changed block identifiers relative to pointed-to blocks at corresponding offsets while bypassing blocks having unchanged volume block numbers.
3. The system of claim 1 or claim 2, further comprising an inode picker for picking out inodes in the retrieved blocks having a predetermined association and that indicate changes between, respectively a first version and a second version of the inodes in the first snapshot and the second snapshot.
4. The system of claim 3, wherein the predetermined association comprises a sub-organization of a volume file system.
5. The system of claim 4, wherein the sub-organization comprises a qtree within the volume file system.
6. The system of claim 4 or claim 5, further comprising an inode worker for operating on queued blocks retrieved by the inode picker and for applying the scanner to derive changes between blocks in a hierarchy of the first version of each of the picked out inodes with respect to a hierarchy of the second version of each of the picked out inodes.
7. The system of claim 6, wherein the inode worker is adapted to collect the changes and forward the changes to a pipeline for packaging the changes in a data stream for transmission to the destination file system.
8. The system of claim 7, further comprising an extensible and file system-independent format that transmits the changes between the source file system and the destination file system.
9. The system of claim 8, wherein the format comprises a plurality of standalone headers including directory information, including deleted files, and basic data file headers, following the standalone headers,

each of the basic data headers being followed by a stream of file data.

10. A method of updating a replica of a logical group of data blocks on a destination with changes in a corresponding logical group on a source, the method comprising:

comparing a first set of block identifiers of a first snapshot of the logical group on the source, which corresponds to the replica, with a second set of block identifiers of a second snapshot of the logical group on the source to identify data blocks in corresponding locations within the logical group that have changed;
transmitting the changed blocks to the destination without transmitting all the blocks of the logical group; and
updating the replica on the destination with the transmitted changed blocks.

11. The method of claim 10, wherein the first snapshot and second snapshot correspond to images of the logical group on the source at respective first and second points in time, and the comparing step selects the corresponding block identifiers of the first and second set of block identifiers that have changed between the first and second points of time.

12. The method of claim 11, wherein the logical group on the source comprises a hierarchical arrangement, and the first and second sets of block identifiers each comprise a hierarchical block index corresponding to the hierarchical arrangement, and the selecting step is performed by a scanner that searches and compares the hierarchical block indexes to identify the changed blocks.

13. The method of claim 11 or claim 12, wherein the transmitting step transmits the changed blocks to a site that is geographically remote from the source.

14. The method of claim 13, wherein the block identifiers of the first and second sets of block identifiers refer to data blocks at corresponding locations within the logical group on the source, and the selecting step selects changed blocks by comparing block identifiers of the first and second sets of blocks identifiers that refer to data blocks at the same locations within the logical group.

15. The method of any one of claims 10 to 14, further comprising picking out inodes, with an inode picker, in the retrieved blocks having a predetermined association and indicating changes between, respectively a first version and a second version of the inodes in the first snapshot and the second snapshot.

16. The method of claim 15, wherein the predetermined association comprises a sub-organization of a volume file system.

17. The method of claim 16, wherein the sub-organization comprises a qtree within the volume file system.

18. A system for receiving a data stream of changes from a snapshot of a logical group of data blocks on a source and updating a replica of the logical group on a destination, the system comprising:

a metadata stage process that reads metadata describing a structure for the logical group on the source and maps references to the logical group on the source with references to the logical group on the destination to identify changed blocks; and

a data stage process that, responsive to the mapped references, populates the logical group on the replica with data blocks from the source that have changed at corresponding offsets in the logical group.

19. A system for identifying changes to a logical group of data blocks on a source and updating a replica of the logical group on a destination, the system comprising:

a metadata stage process including the steps of (i) reading metadata describing a structure for the logical group on the source; the metadata including a first set of references to the data blocks of the logical group on the source; and (ii) generating a replica of the read metadata, including the step of mapping the first set of references to a second set of references to data blocks of the logical group on the destination; and

a data stage process responsive to the metadata stage process including the step of populating the replica with data blocks referenced by the second set of references.

20. A method of updating a replica on a destination, the method comprising:

reading, from changed data of the snapshot, identifiers related to deleted and modified logical groups of data on the replica and placing the deleted and modified logical groups in a temporary store separate from a main store of the replica;
creating a set of references in the main store to the deleted and modified logical groups in the temporary store; and
after the creating step, deallocating the temporary store while maintaining the references in

the main store to the deleted and modified logical groups of data.

21. A carrier medium carrying computer readable code for controlling a computer to carry out the method of any one of claims 10 to 17 or 20 5

10

15

20

25

30

35

40

45

50

55

22

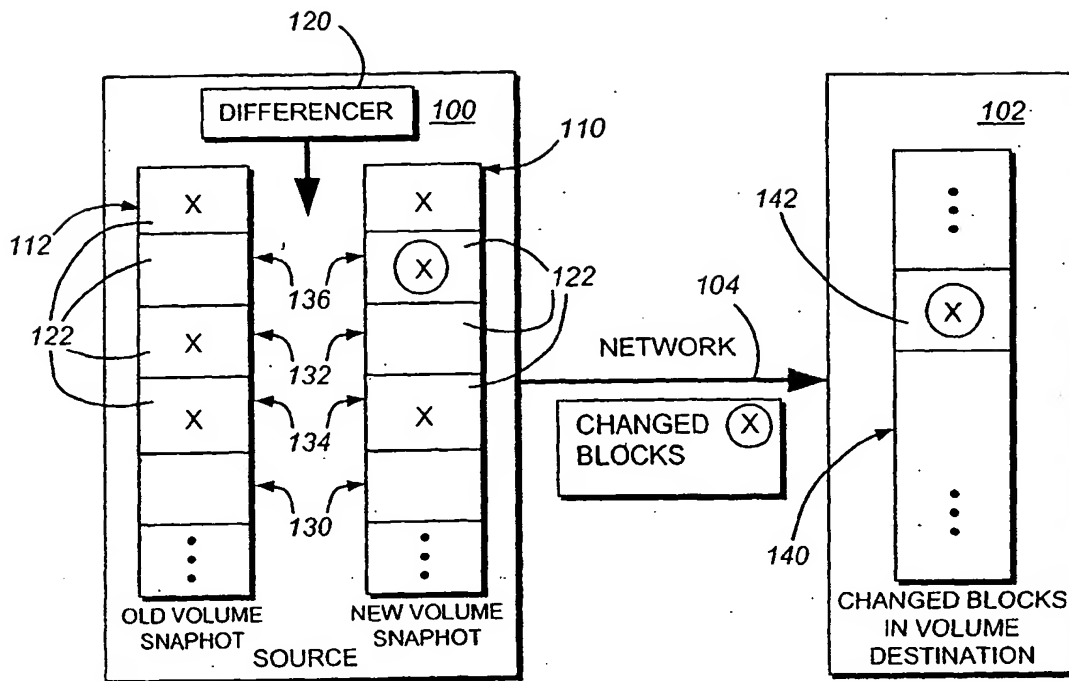


FIG. 1
(PRIOR ART)

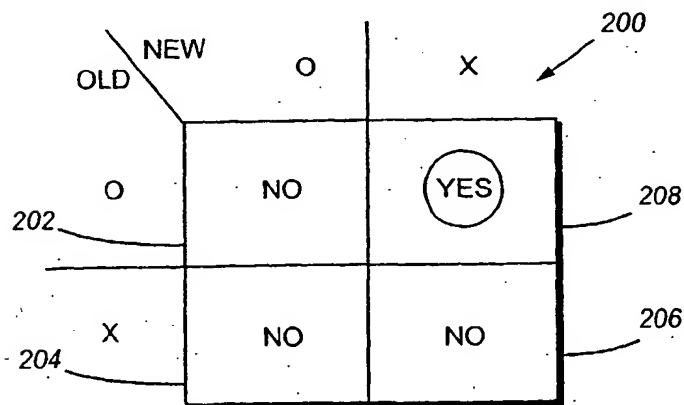


FIG. 2
(PRIOR ART)

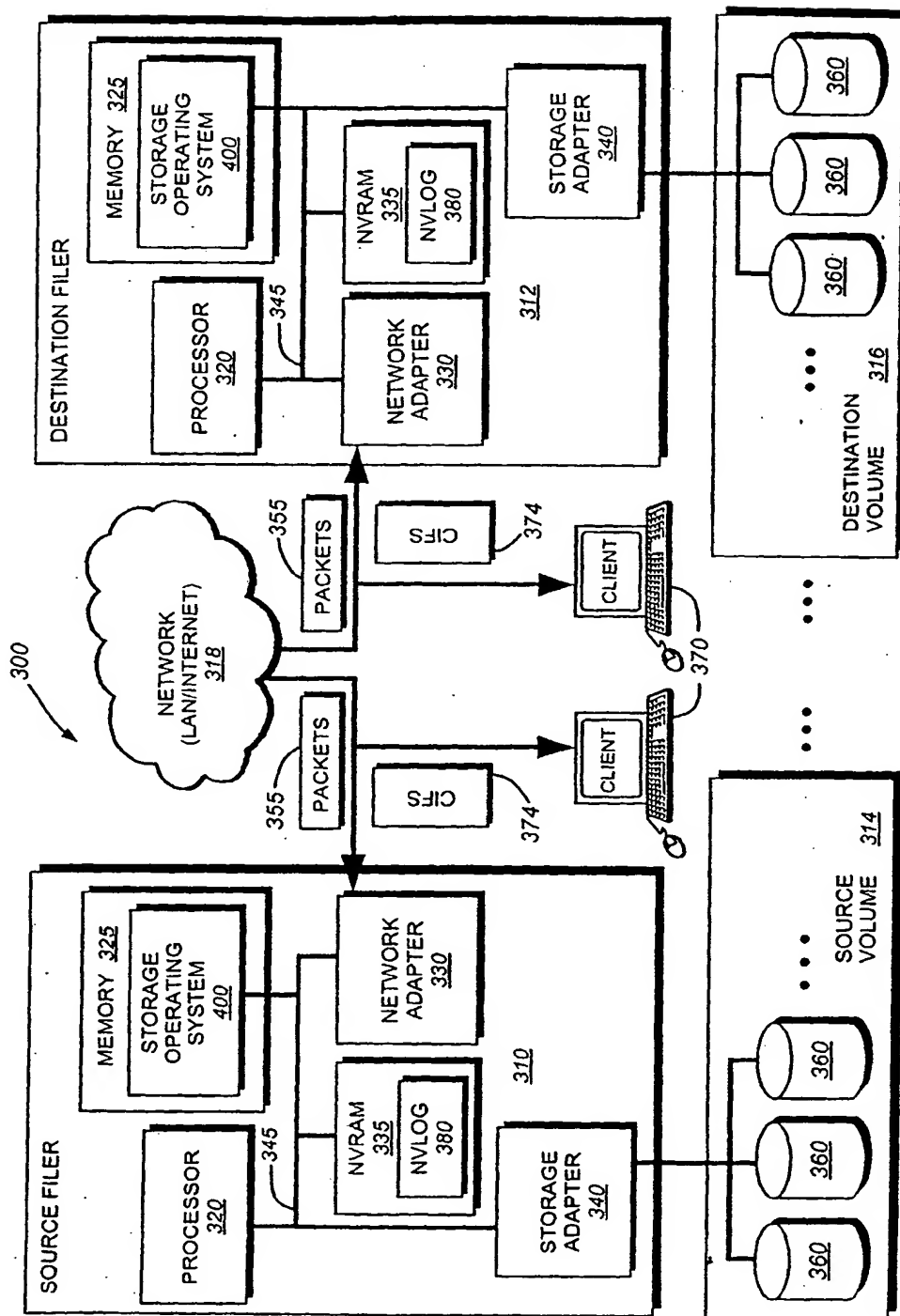
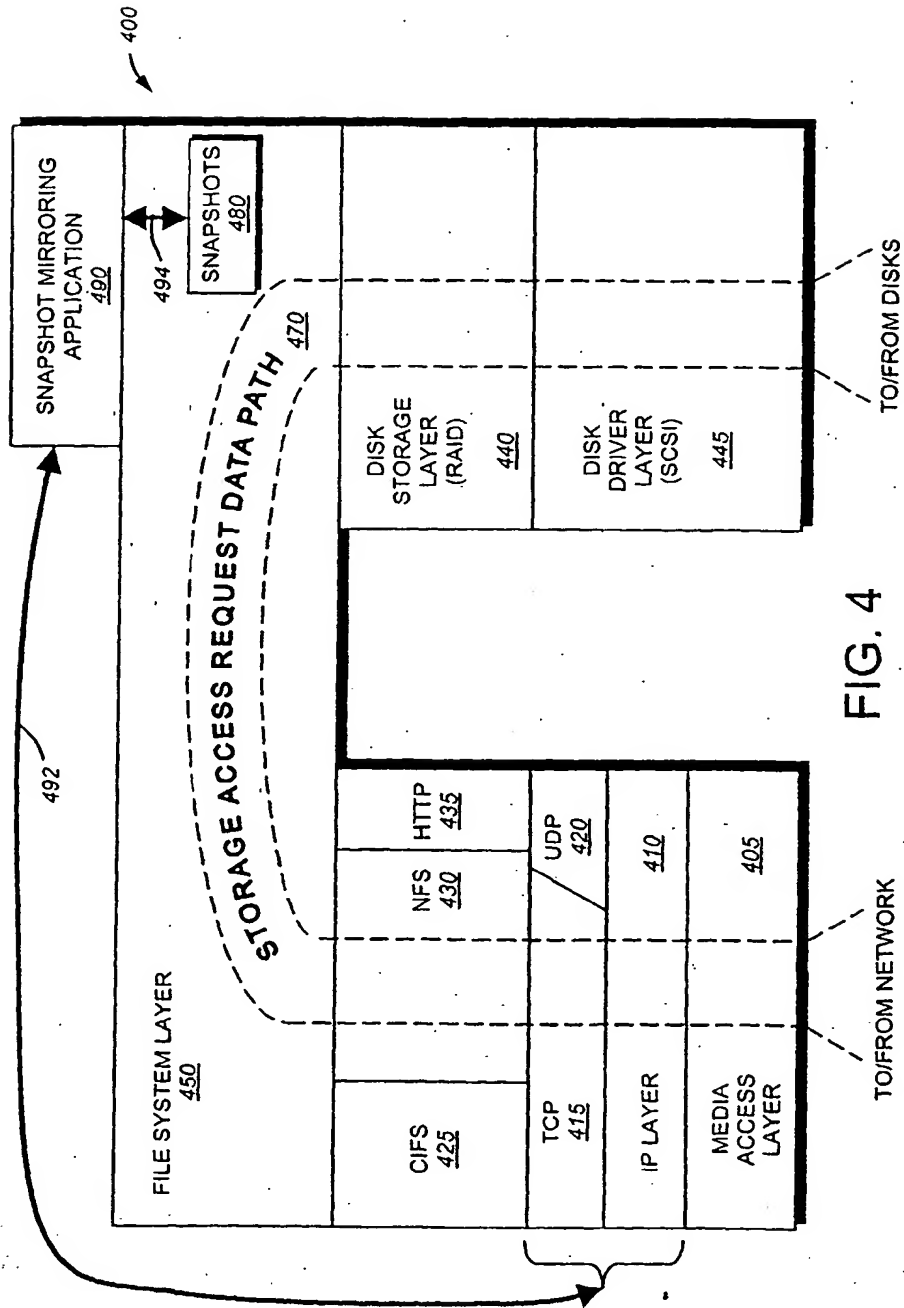


FIG. 3



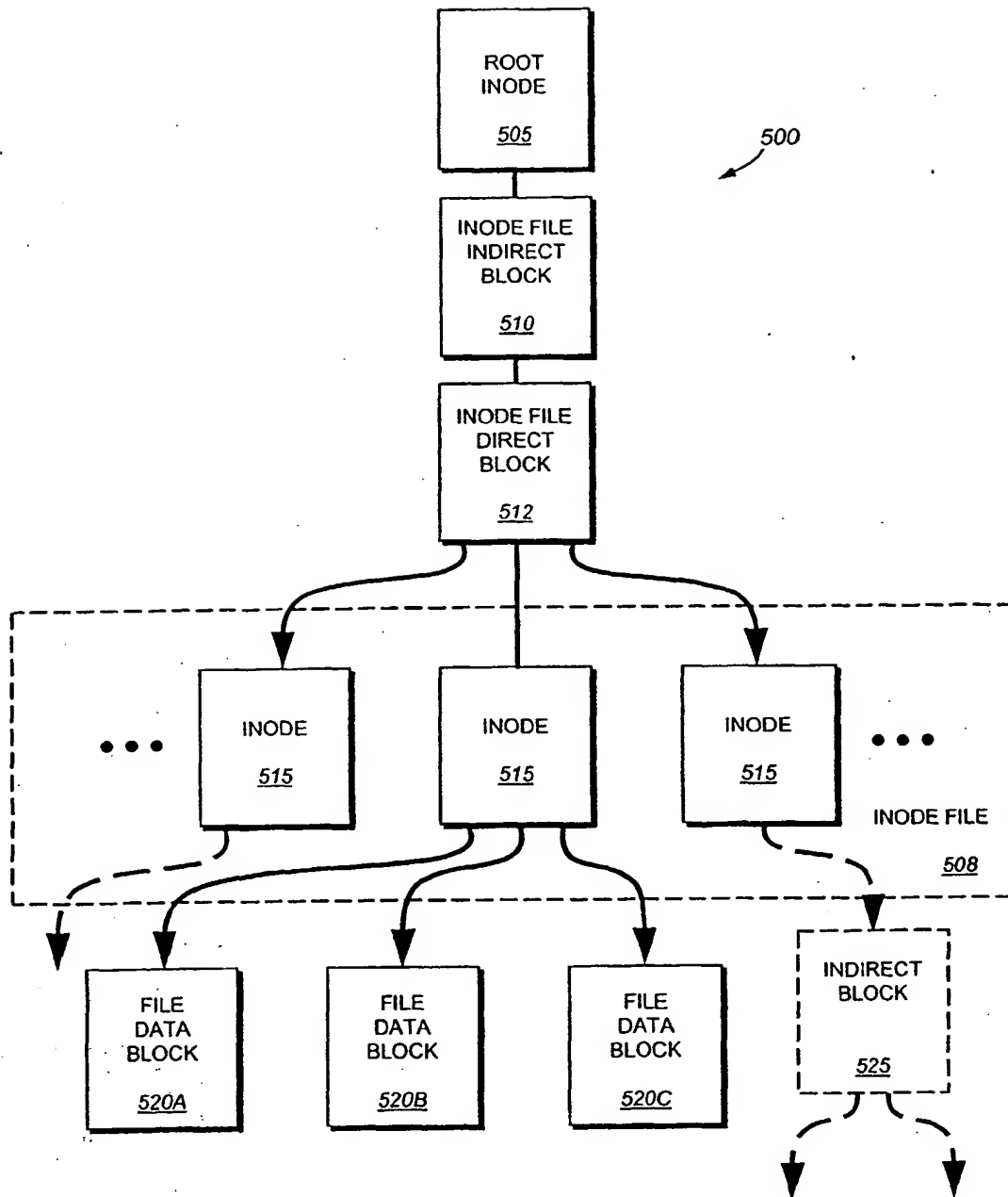


FIG. 5

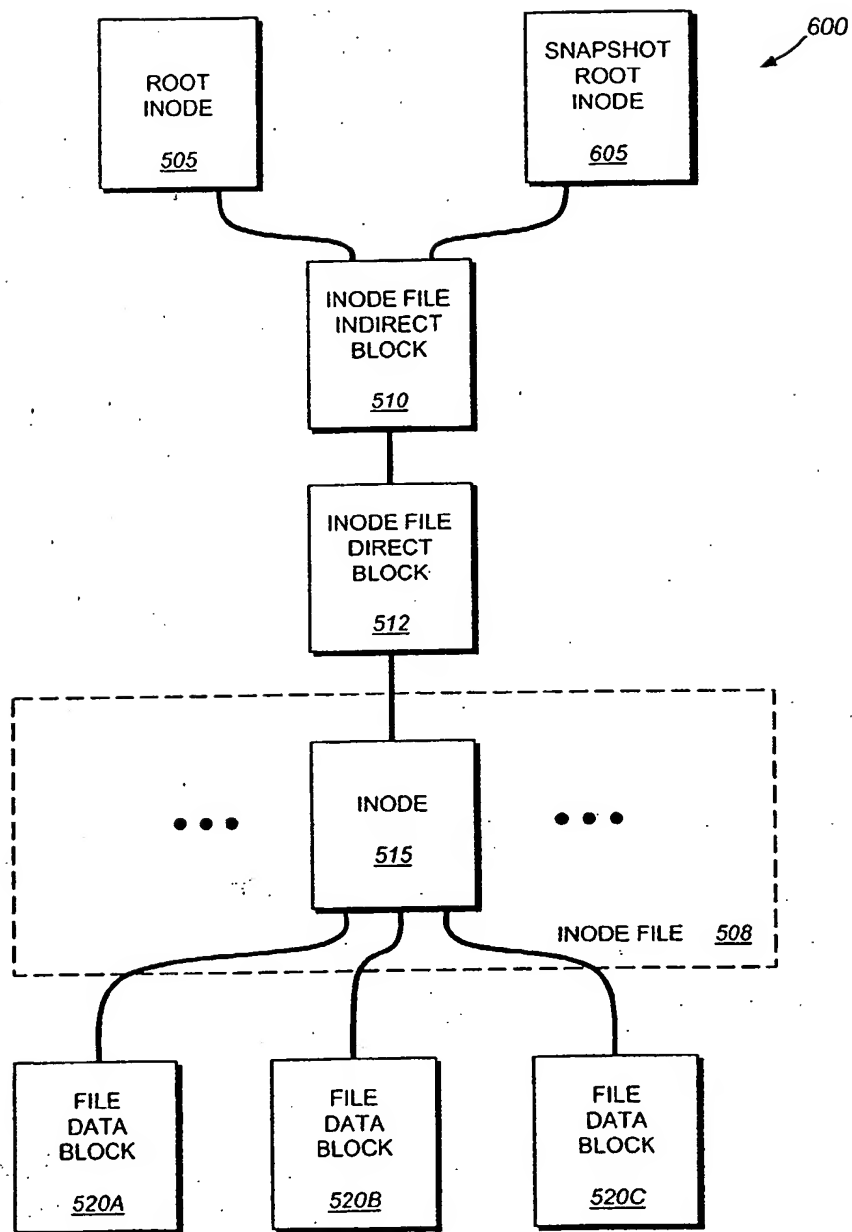


FIG. 6

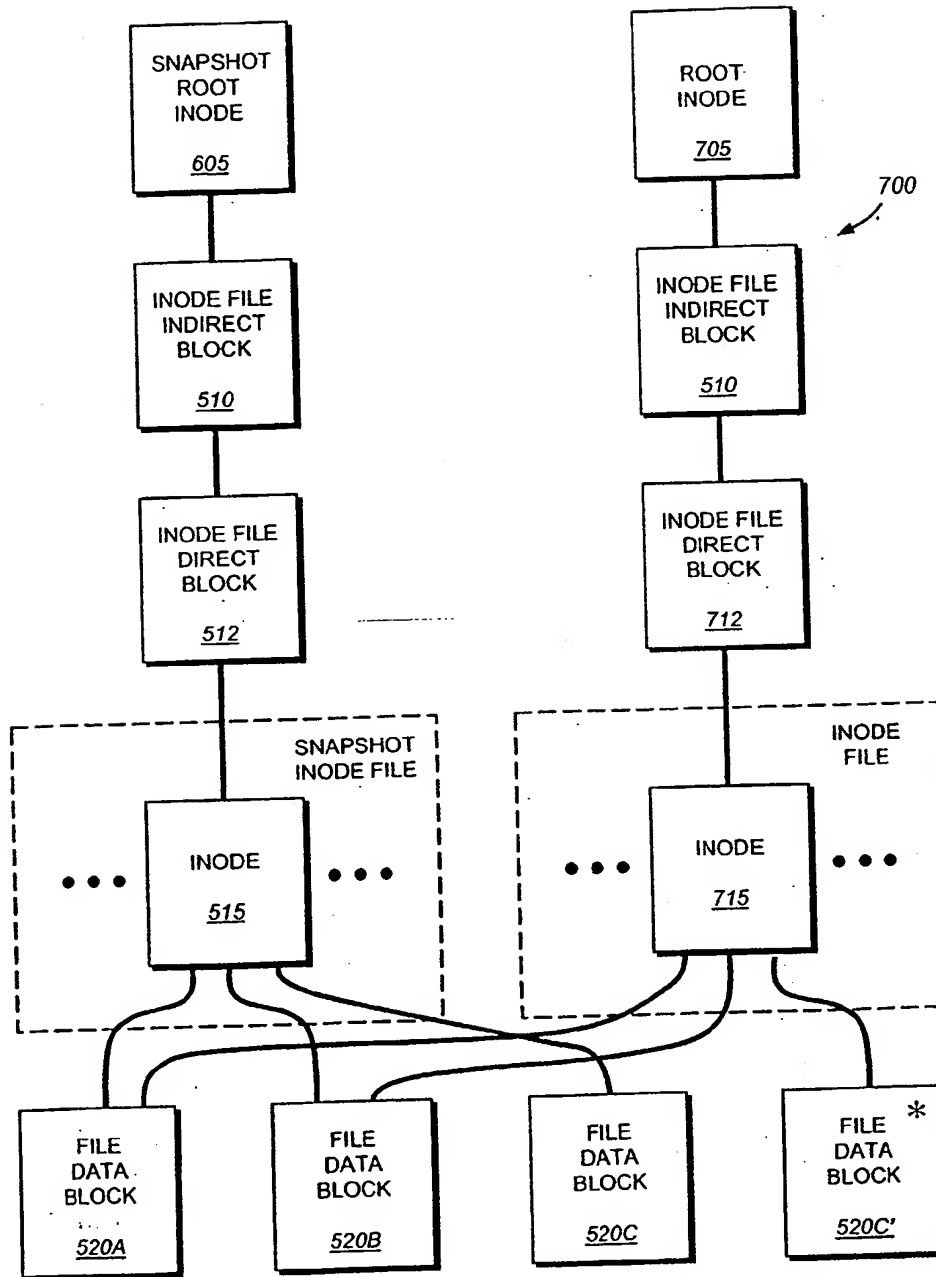


FIG. 7

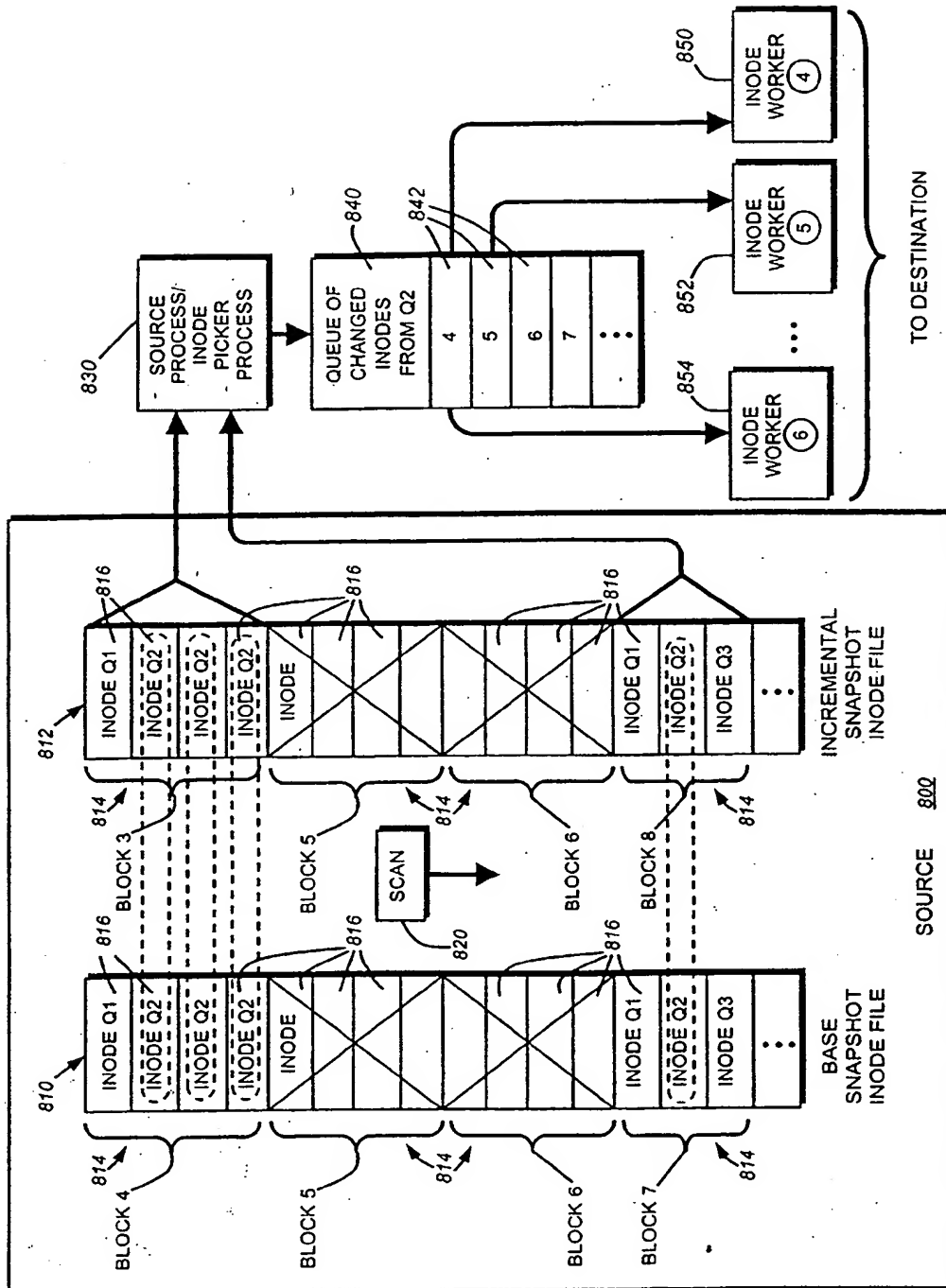


FIG. 8

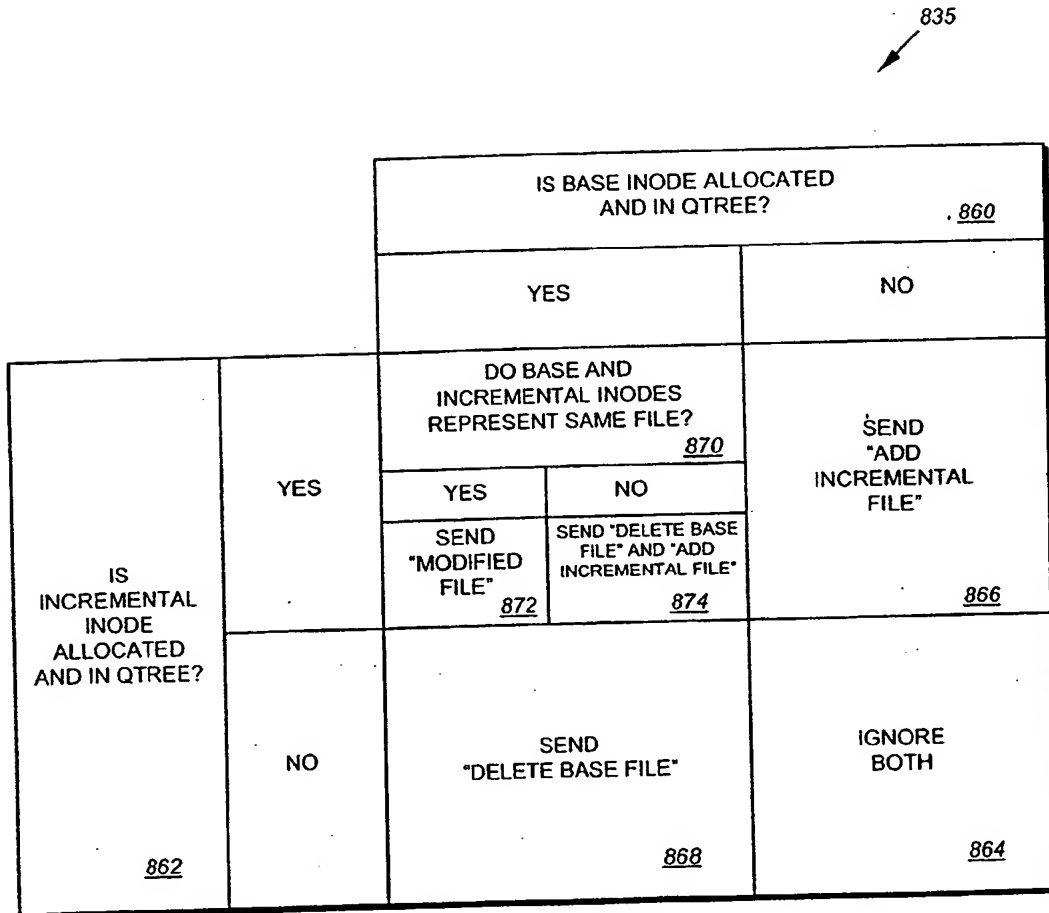


FIG. 8A

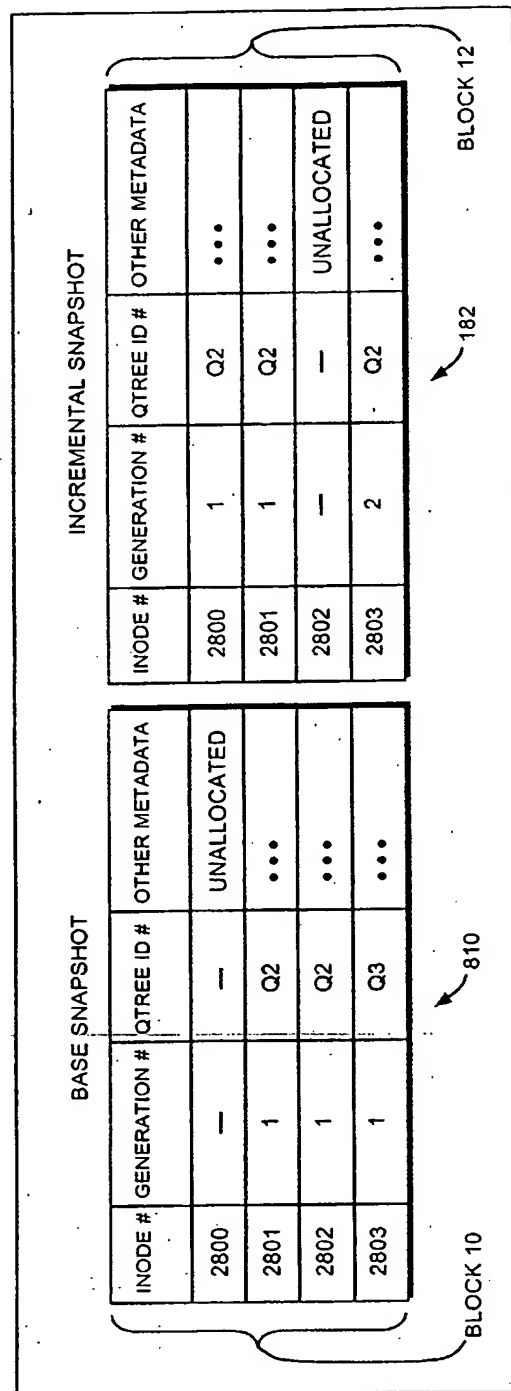


FIG. 8B

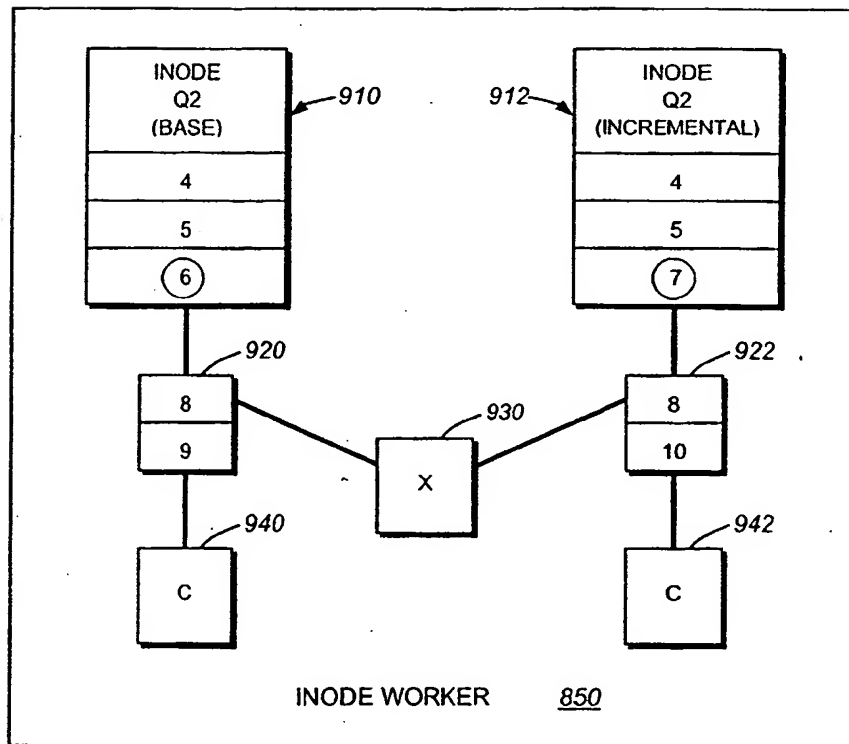


FIG. 9

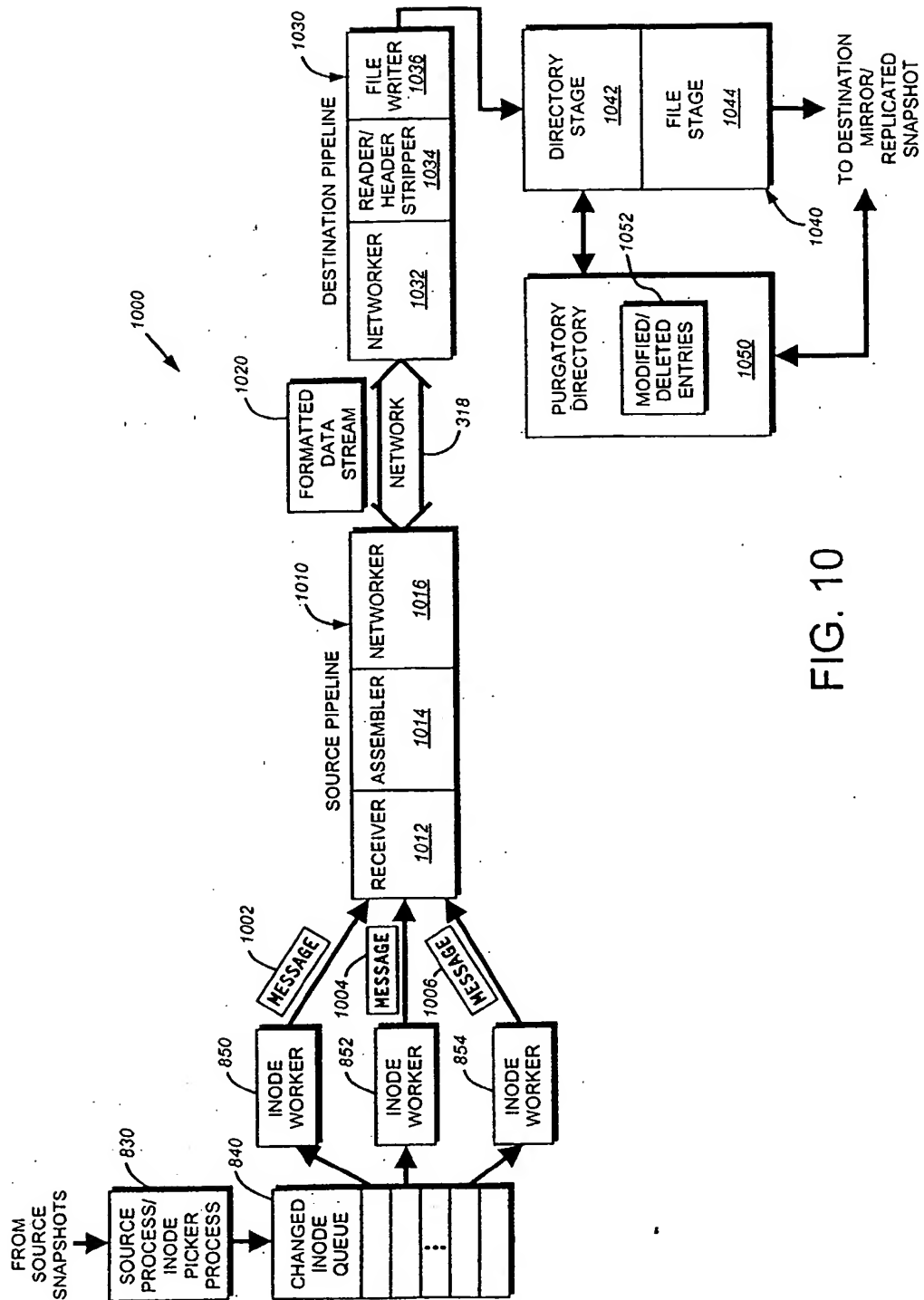


FIG. 10

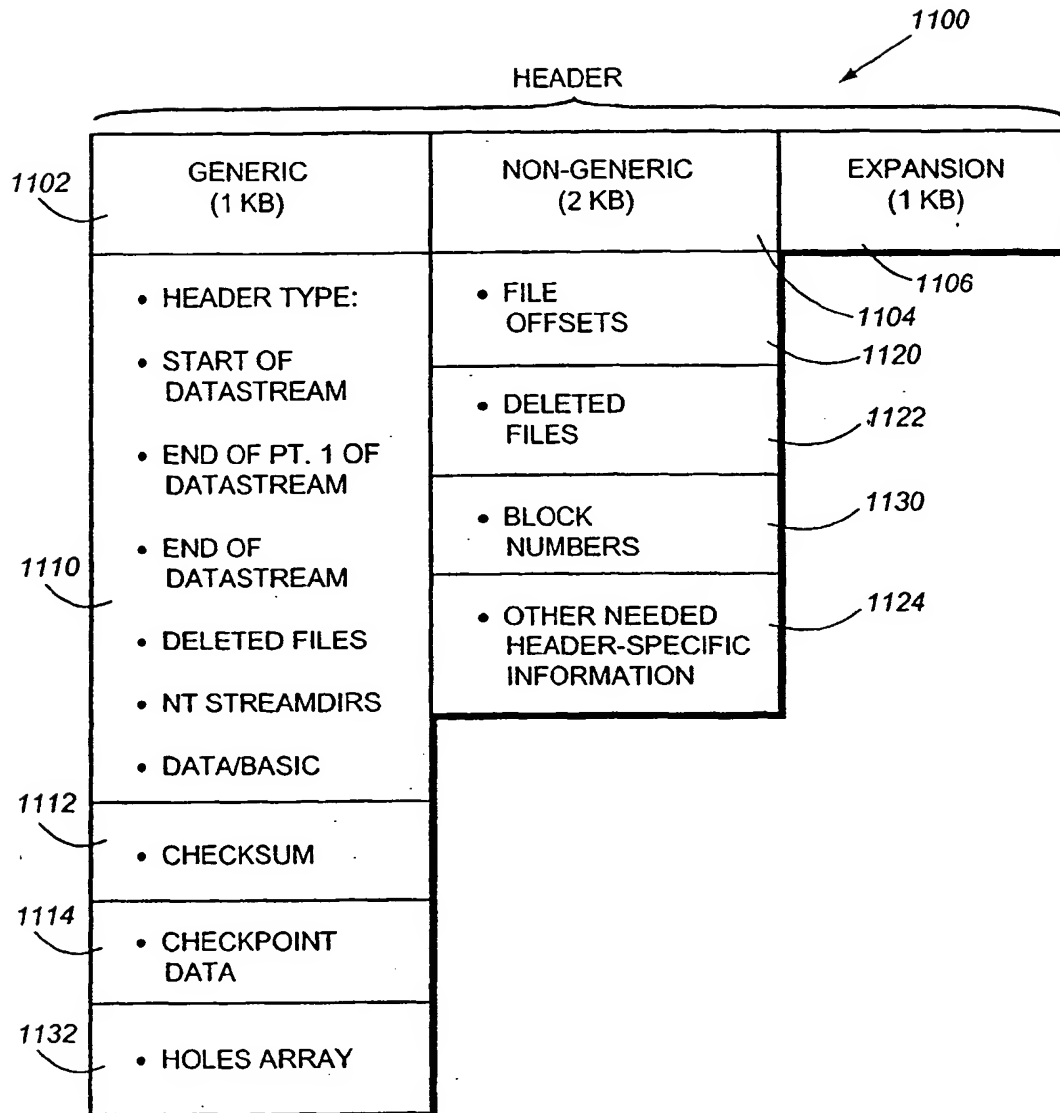


FIG. 11

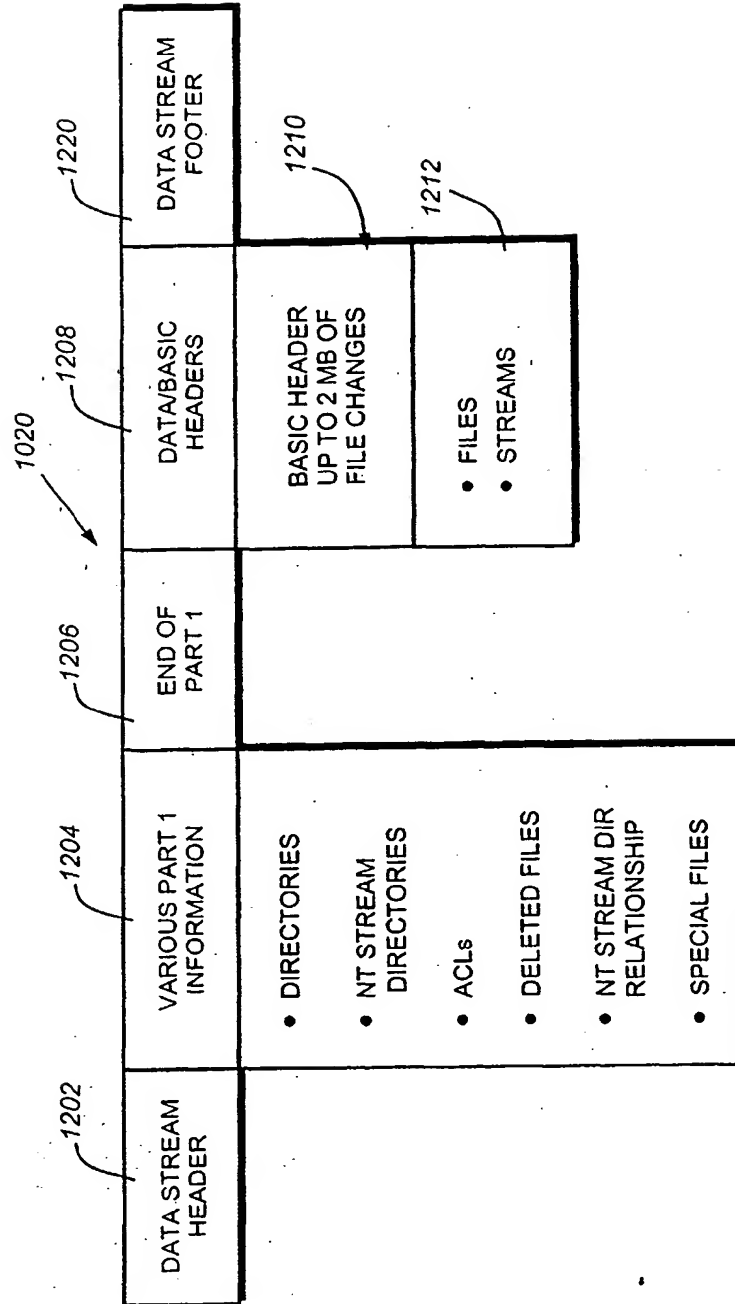
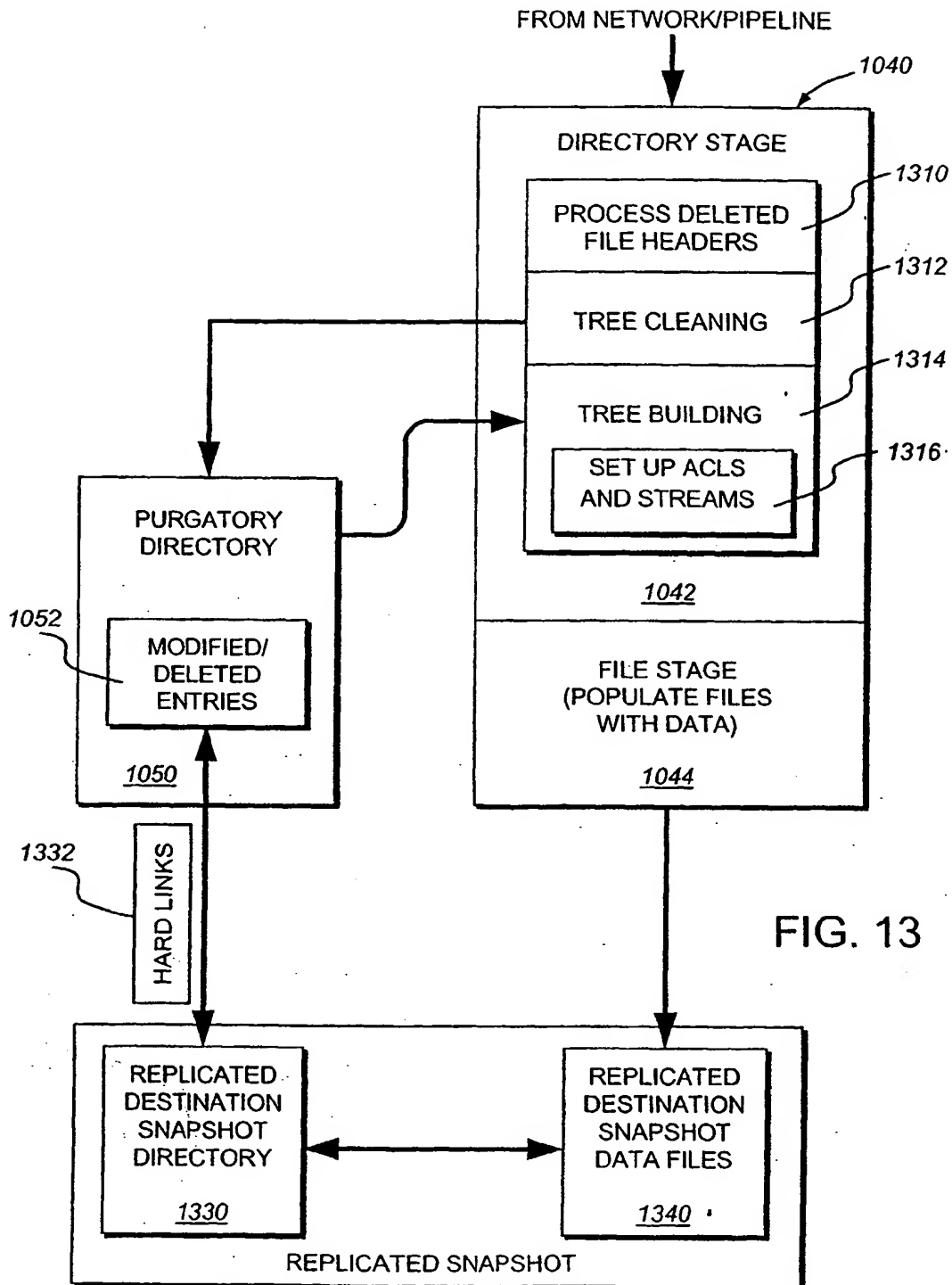


FIG. 12



1400

1404	SOURCE INODE	INODE 877	INODE 878	...
1406	DESTINATION INODE	INODE 9912	INODE 10100	...
1408	SOURCE GENERATION #	3	2	...
1410	DESTINATION GENERATION #	3	3	...

1402

FIG. 14

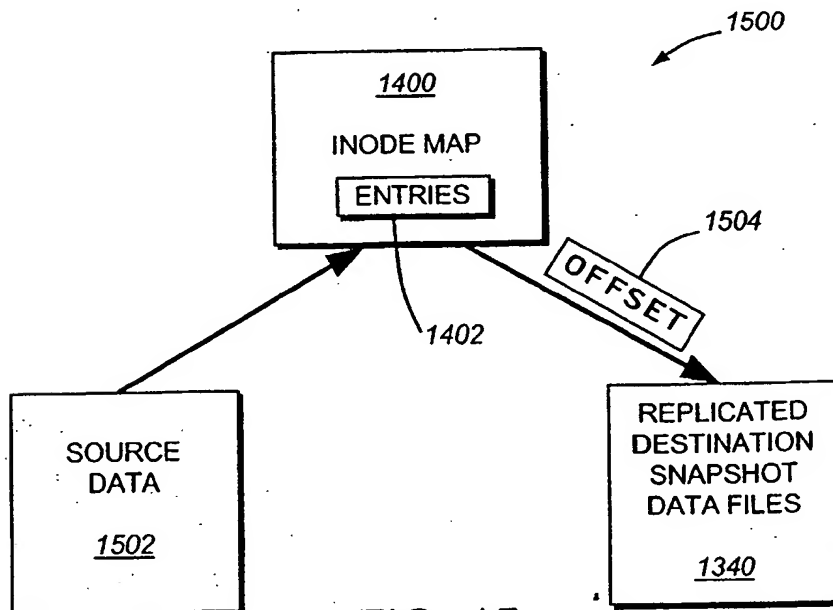


FIG. 15

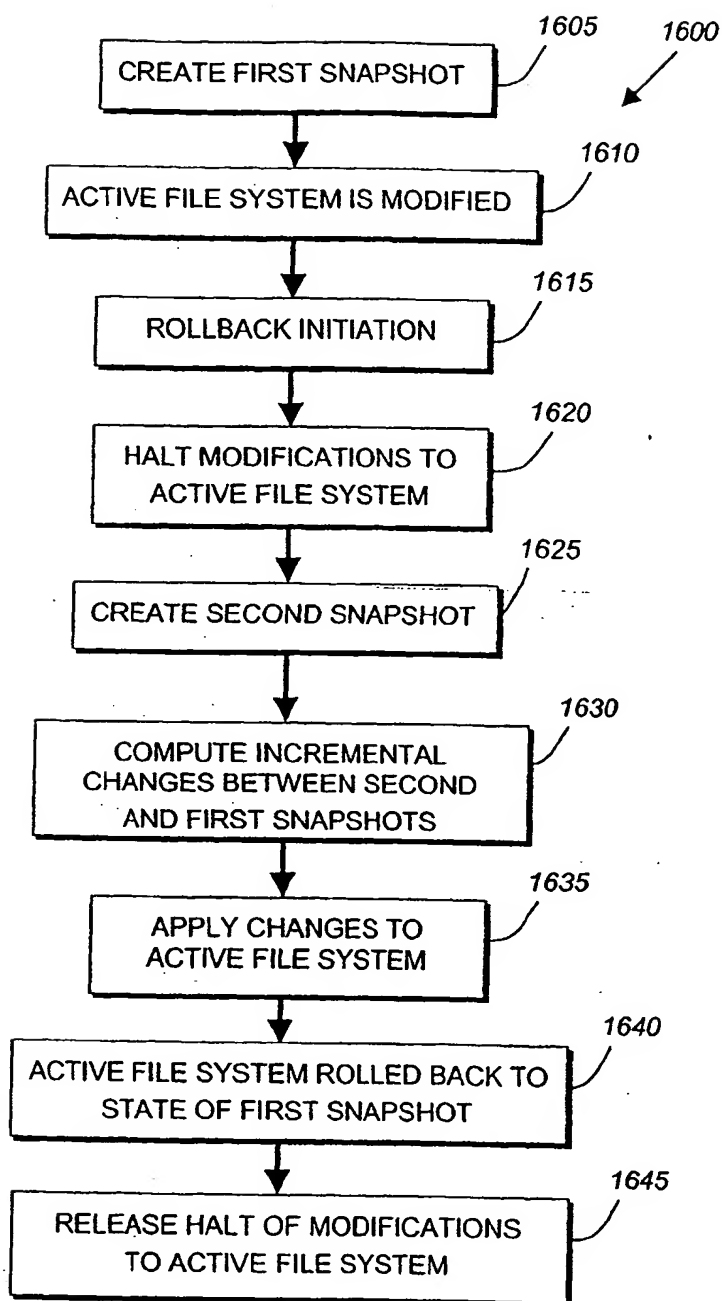


FIG. 16

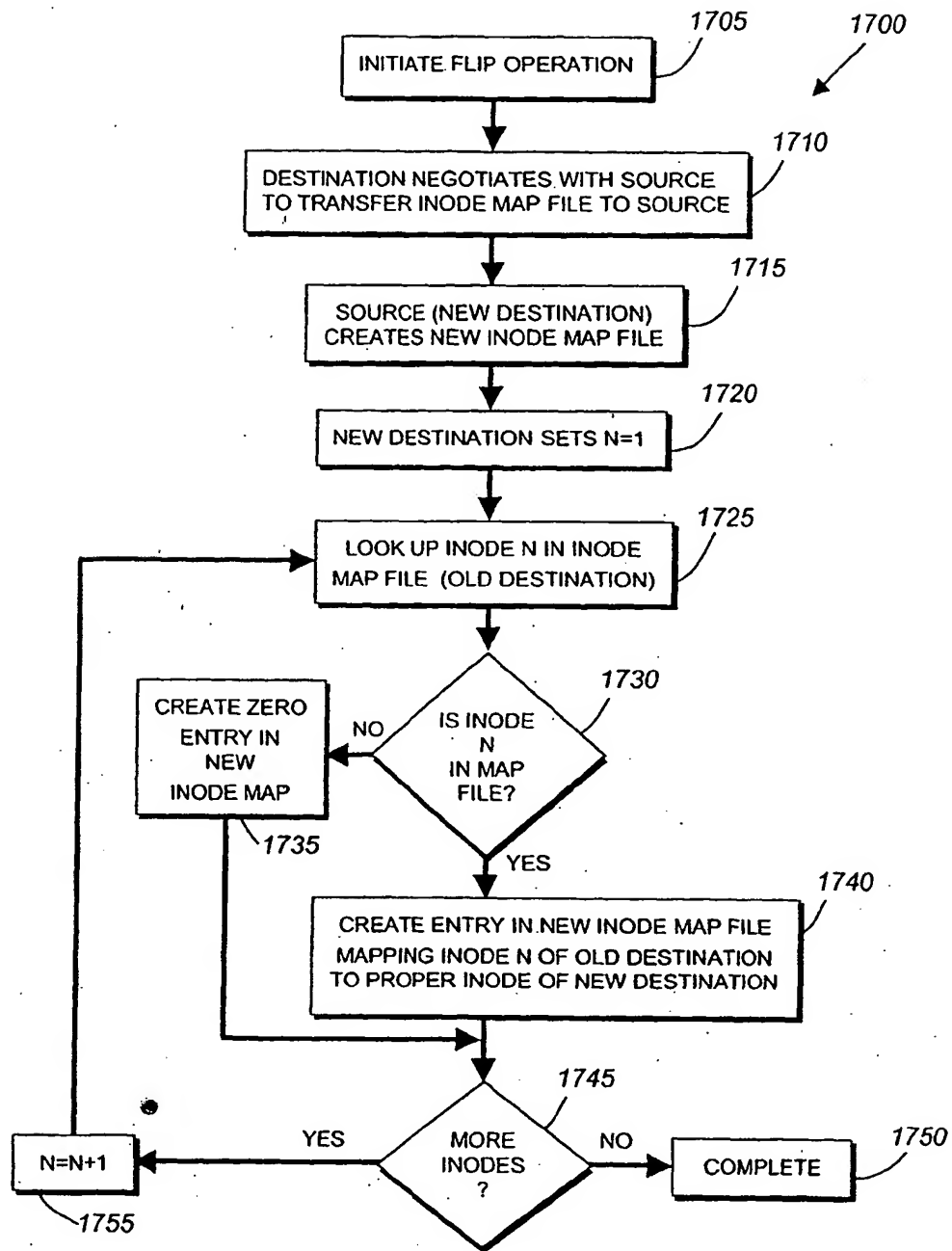


FIG. 17

1800

...	SOURCE INODE 72	SOURCE INODE 83	SOURCE INODE 190	...
...	DESTINATION INODE 605	DESTINATION INODE 328	DESTINATION INODE 150	...
...	1	2	2	...
...	2	2	2	...

1802 1804 1806

1404 1406 1408 1410

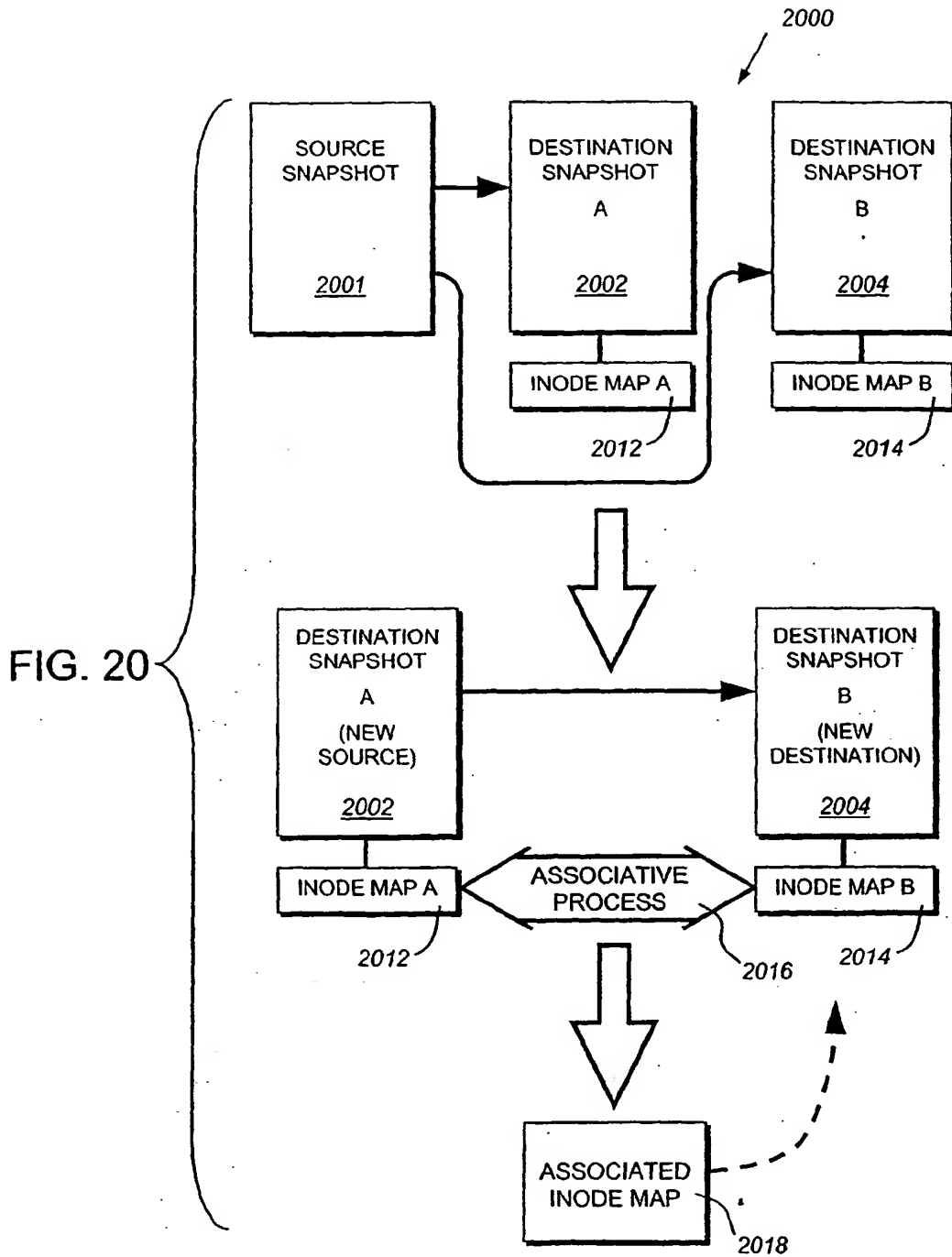
FIG. 18

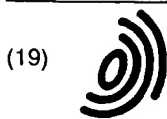
1900

...	NEW SOURCE INODE 150	NEW SOURCE INODES 151-327 ...	NEW SOURCE INODE 328	NEW SOURCE INODES 329-604 ...	NEW SOURCE INODE 605	NEW SOURCE INODE 606	...
...	NEW DESTINATION INODE 190	...	NEW DESTINATION INODE 83	...	NEW DESTINATION INODE 72	0	...
...	2	...	2	...	2	-	...
...	2	...	2	...	1	-	...

1902 1904 1906 1908 1910 1914 1912 1918 1916 1930

FIG. 19





Europäisches Patentamt
European Patent Office
Office européen des brevets



(11) EP 1 349 088 A3

(12) EUROPEAN PATENT APPLICATION

(88) Date of publication A3:
20.04.2005 Bulletin 2005/16

(51) Int Cl.7: G06F 17/30, G06F 11/14,
G06F 11/20

(43) Date of publication A2:
01.10.2003 Bulletin 2003/40

(21) Application number: 03251702.1

(22) Date of filing: 19.03.2003

(84) Designated Contracting States:
AT BE BG CH CY CZ DE DK EE ES FI FR GB GR
HU IE IT LI LU MC NL PT RO SE SI SK TR
Designated Extension States:
AL LT LV MK

- Manley, Stephen L.
London SW5 9JZ (GB)
- Owara, Shane S.
Mountain View, California 94040 (US)
- Kleiman, Steven R.
Sunnyvale, California 94089 (US)

(30) Priority: 19.03.2002 US 100950
19.03.2002 US 100967

(71) Applicant: Network Appliance, Inc.
Sunnyvale, California 94089 (US)

(74) Representative: Collins, John David et al
Marks & Clerk
90 Long Acre
London WC2E 9RA (GB)

(72) Inventors:
• Federwisch, Michael L.
San Jose, California 95125 (US)

(54) System and method for determining and transmitting changes in snapshots

(57) A system and method for remote asynchronous replication or mirroring of changes in a source file system snapshot in a destination replica file system using a scan (via a scanner) of the blocks that make up two versions of a snapshot of the source file system, which identifies changed blocks in the respective snapshot files based upon differences in volume block numbers identified in a scan of the logical file block index of each snapshot. Trees of blocks associated with the files are traversed, bypassing unchanged pointers between versions and walking down to identify the changes in the hierarchy of the tree. These changes are transmitted to the destination mirror or replicated snapshot. This technique allows regular files, directories, inodes and any other hierarchical structure to be efficiently scanned to determine differences between versions thereof. The changes in the files and directories are transmitted over the network for update of the replicated destination snapshot in an asynchronous (lazy write) manner. The changes are described in an extensible, system-independent data stream format layered under a network transport protocol. At the destination, source changes are used to update the destination snapshot. Any deleted or modified inodes already on the destination are moved to a temporary or "purgatory" directory and, if re-used, are relinked to the rebuilt replicated snapshot di-

rectory. The source file system snapshots can be representative of a volume sub-organization, such as a qtree.

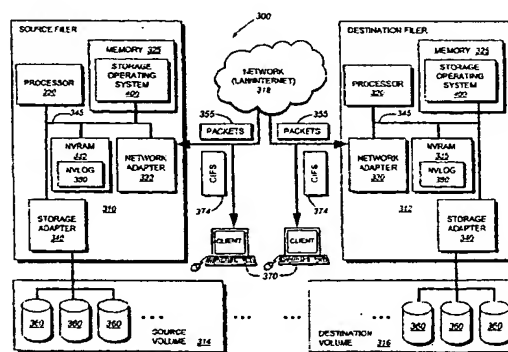


FIG. 3

EP 1 349 088 A3



European Patent
Office

EUROPEAN SEARCH REPORT

Application Number
EP 03 25 1702

DOCUMENTS CONSIDERED TO BE RELEVANT			
Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.7)
X	WO 00/07104 A (NETWORK APPLIANCE INC) 10 February 2000 (2000-02-10)	1,10,11, 13,14,21	G06F17/30 G06F11/14
Y	* abstract *	2-9,12, 15-17	G06F11/20
	* page 9, line 19 - page 11, column 30 * * page 19, line 21 - page 21, line 22 * -----		
Y	EP 1 003 103 A (NETWORK APPLIANCE INC) 24 May 2000 (2000-05-24)	2-9,12, 15-17	
A	* abstract; figures 18,22 *	1,10,11, 13,14,21	
	* paragraphs [0089] - [0091], [0098], [0099] *		
A	"STORAGE MANAGEMENT MECHANISM FOR MANAGING SNAPSHOT PAGES" IBM TECHNICAL DISCLOSURE BULLETIN, IBM CORP. NEW YORK, US, vol. 35, no. 4B, 1 September 1992 (1992-09-01), pages 26-29, XP000313841 ISSN: 0018-8689 * the whole document *	1,10, 18-21	

			TECHNICAL FIELDS SEARCHED (Int.Cl.7)
			G06F

-The present search report has been drawn up for all claims			
Place of search Munich		Date of completion of the search 16 December 2004	Examiner Leuridan, K
CATEGORY OF CITED DOCUMENTS		Y : theory or principle underlying the invention E : earlier patent document, but published on, or after the filing date D : document cited in the application L : document cited for other reasons ----- & : member of the same patent family, corresponding document	
X : particularly relevant if taken alone Y : particularly relevant if combined with another document of the same category A : technological background O : non-written disclosure P : intermediate document			

EPO FORM 1503 03 82 (P04C01)



European Patent
Office

Application Number

EP 03 25 1702

CLAIMS INCURRING FEES

The present European patent application comprised at the time of filing more than ten claims.

- ☐ Only part of the claims have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims and for those claims for which claims fees have been paid, namely claim(s):
- ☐ No claims fees have been paid within the prescribed time limit. The present European search report has been drawn up for the first ten claims.

LACK OF UNITY OF INVENTION

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions, namely:

see sheet B

- ☐ All further search fees have been paid within the fixed time limit. The present European search report has been drawn up for all claims.
- ☐ As all searchable claims could be searched without effort justifying an additional fee, the Search Division did not invite payment of any additional fee.
- ☐ Only part of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the inventions in respect of which search fees have been paid, namely claims:
- ☒ None of the further search fees have been paid within the fixed time limit. The present European search report has been drawn up for those parts of the European patent application which relate to the invention first mentioned in the claims, namely claims:

1-17, 21



European Patent
Office

LACK OF UNITY OF INVENTION
SHEET B

Application Number
EP 03 25 1702

The Search Division considers that the present European patent application does not comply with the requirements of unity of invention and relates to several inventions or groups of inventions, namely:

1. claims: 1-17,21

A method of comparing 2 snapshots on a source

2. claims: 18,19

A system for updating a replica

3. claim: 20

A method for maintaining consistency while updating a replica.

**ANNEX TO THE EUROPEAN SEARCH REPORT
ON EUROPEAN PATENT APPLICATION NO.**

EP 03 25 1702

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report.
The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

16-12-2004

Patent document cited in search report		Publication date		Patent family member(s)	Publication date
WO 0007104	A	10-02-2000	US	2001044807 A1	22-11-2001
			AT	278216 T	15-10-2004
			DE	69920713 D1	04-11-2004
			EP	1099165 A1	16-05-2001
			WO	0007104 A1	10-02-2000
			US	6574591 B1	03-06-2003
			US	2003217082 A1	20-11-2003
			US	2002049718 A1	25-04-2002

EP 1003103	A	24-05-2000	EP	1003103 A2	24-05-2000
			AT	195825 T	15-09-2000
			DE	69425658 D1	28-09-2000
			DE	69425658 T2	19-04-2001
			DK	702815 T3	18-12-2000
			EP	0702815 A1	27-03-1996
			HK	1013697 A1	17-08-2001
			JP	8511367 T	26-11-1996
			WO	9429807 A1	22-12-1994
			US	6289356 B1	11-09-2001
			US	5819292 A	06-10-1998
			US	2004260673 A1	23-12-2004
			US	2002049718 A1	25-04-2002
			US	2002091670 A1	11-07-2002

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

This Page Blank (uspto)

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.

This Page Blank (uspto)